Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts

Jana Diesner

September 2012

CMU-ISR-12-101

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Kathleen M. Carley (CMU, ISR), chair
William W. Cohen (CMU, LTI)
Carolyn P. Rosé (CMU, LTI)
Jeffrey Johnson (East Carolina University)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

This work was supported by the National Science Foundation (NSF) IGERT 9972762, the Army Research Institute (ARI) W91WAW07C0063, the Army Research Laboratory (ARL/CTA) DAAD19-01- 2-0009, the Air Force Office of Scientific Research (AFOSR) MURI FA9550-05-1-0388, the Office of Naval Research (ONR) MURI N00014-08-11186, and a Siebel Scholarship. Additional support was provided by CASOS, the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, ARI, ARL, AFOSR, ONR, or the United States Government.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **SEP 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**
4. TITLE AND SUBTITLE **Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Institute for Software Research,Pittsburgh,PA,15213**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR’S ACRONYM(S)
		11. SPONSOR/MONITOR’S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES		

14. ABSTRACT

This thesis is motivated by the need for scalable and reliable methods and technologies that support the construction of network data based on information from text data. Ultimately, the resulting data can be used for answering substantive and graph-theoretical questions about sociotechnical networks. One main limitation with constructing network data from text data is that the validation of the resulting network data can be hard to infeasible, e.g. in the cases of covert, historical and largescale networks. This thesis addresses this problem by identifying the impact of coding choices that must be made when extracting network data from text data on the structure of networks and network analysis results. My findings suggest that conducting reference resolution on text data can alter the identity and weight of 76% of the nodes and 23% of the links, and can cause major changes in the value of commonly used network metrics. Also, performing reference resolution prior to relation extraction leads to the retrieval of completely different sets of key entities in comparison to not applying this pre-processing technique. Based on the outcome of the presented experiments, I recommend strategies for avoiding or mitigating the identified issues in practical applications. When extracting socio-technical networks from texts, the set of relevant node classes might go beyond the classes that are typically supported by tools for named entity extraction. I address this lack of technology by developing an entity extractor that combines an ontology for sociotechnical networks that originates from the social sciences, is theoretically grounded and has been empirically validated in prior work, with a supervised machine learning technique that is based on probabilistic graphical models. This thesis does not stop at showing that the resulting prediction models achieve state of the art accuracy rates, but I also describe the process of integrating these models into an existing and publically available end-user product. As a result users can apply these models to new text data in a convenient fashion. While a plethora of methods for building network data from information explicitly or implicitly contained in text data exists, there is a lack of research on how the resulting networks compare with respect to their structure and properties. This also applies to networks that can be extracted by using the aforementioned entity extractor as part of the relation extraction process. I address this knowledge gap by comparing the networks extracted by using this process to network data built with three alternative methods: text coding based on thesauri that associate text terms with node classes, the construction of network data from meta-data on texts, such as key words and

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**337**	

Keywords: socio-technical networks, semantic networks, information networks, entity extraction, relation extraction, reference resolution, co-occurrence based network construction, evaluation and comparison of supervised and unsupverised learning for text coding, accuracy assessment, network clustering, grouping in networks

Abstract

This thesis is motivated by the need for scalable and reliable methods and technologies that support the construction of network data based on information from text data. Ultimately, the resulting data can be used for answering substantive and graph-theoretical questions about socio-technical networks.

One main limitation with constructing network data from text data is that the validation of the resulting network data can be hard to infeasible, e.g. in the cases of covert, historical and large-scale networks. This thesis addresses this problem by identifying the impact of coding choices that must be made when extracting network data from text data on the structure of networks and network analysis results. My findings suggest that conducting reference resolution on text data can alter the identity and weight of 76% of the nodes and 23% of the links, and can cause major changes in the value of commonly used network metrics. Also, performing reference resolution prior to relation extraction leads to the retrieval of completely different sets of key entities in comparison to not applying this pre-processing technique. Based on the outcome of the presented experiments, I recommend strategies for avoiding or mitigating the identified issues in practical applications.

When extracting socio-technical networks from texts, the set of relevant node classes might go beyond the classes that are typically supported by tools for named entity extraction. I address this lack of technology by developing an entity extractor that combines an ontology for socio-technical networks that originates from the social sciences, is theoretically grounded and has been empirically validated in prior work, with a supervised machine learning technique that is based on probabilistic graphical models. This thesis does not stop at showing that the resulting prediction models achieve state of the art accuracy rates, but I also describe the process of integrating these models into an existing and publically available end-user product. As a result, users can apply these models to new text data in a convenient fashion.

While a plethora of methods for building network data from information explicitly or implicitly contained in text data exists, there is a lack of research on how the resulting networks compare with respect to their structure and properties. This also applies to networks that can be extracted by using the aforementioned entity extractor as part of the relation extraction process. I address this knowledge gap by comparing the networks extracted by using this process to network data built with three alternative methods: text coding based on thesauri that associate text terms with node classes, the construction of network data from meta-data on texts, such as key words and index terms, and building network data in collaboration with subject matter experts. The outcomes of these comparative analyses suggest that thesauri generated with the entity extractor developed for this thesis need adjustments with respect to particular categories and types of errors. I am providing tools and strategies to assist with these refinements. My results also show that once these changes have been made and in contrast to manually constructed thesauri, the prediction models generalize with acceptable accuracy to other domains (news wire data, scientific writing, emails) and writing styles (formal, casual). The comparisons of networks constructed with different methods show that ground truth data built by subject matter experts are hardly resembled by any automated method that analyzes text bodies, and even less so by exploiting existing meta-data from text corpora. Thus, aiming to reconstruct social networks from text data leads to largely incomplete networks. Synthesizing the findings from this work, I outline which types of information on socio-technical networks are best captured by what network data construction method, and how to best combine these methods in order to gain a more comprehensive view on a network.

When both, text data and relational data, are available as a source of information on a network, people have previously integrated these data by enhancing social networks with content nodes that represent salient terms from the text data. I present a methodological advancement to this technique and test its performance on the datasets used for the previously mentioned evaluation studies. By using this approach, multiple types of behavioral data, namely interactions between people as well as their language use, can be taken into account. I conclude that extracting content nodes from groups of structurally equivalent agents can be an appropriate strategy for enabling the comparison of the content that people produce, perceive or disseminate. These equivalence classes can represent a variety of social roles and social positions that network members occupy. At the same time, extracting content nodes from groups of structurally coherent agents can be suitable for enabling the enhancement of social networks with content nodes. The results from applying the latter approach to text data include a comparison of the outcome of topic modeling; an efficient and unsupervised information extraction technique, to the outcomes of alternative methods, including entity extraction based on supervised machine learning. My findings suggest that key entities from meta-data knowledge networks might serve as proper labels for unlabeled topics. Also, unsupervised and supervised learning leads to the retrieval of similar entities as highly likely members of highly likely topics, and key nodes from text-based knowledge networks, respectively.

In summary, the contributions made with this thesis help people to collect, manage and analyze rich network data at any scale. This is a precondition for asking substantive and graph-theoretical questions, testing hypotheses, and advancing theories about networks. This thesis uses an interdisciplinary and computationally rigorous approach to work towards this goal; thereby advancing the intersection of network analysis, natural language processing and computing.

Acknowledgement

First and foremost, I would like to thank my advisor, Kathleen M. Carley, for her guidance and support. Kathleen was instrumental in showing me how to bring social science and computer science together in order to solve relevant, real-world problems through empirical research. I am grateful to Kathleen for continually pushing me to tie my research findings to the practical implications of our work, and for training me how to always keep the bigger picture with our research projects in mind. Kathleen taught me how to become an effective writer and how to stay focused on making our research outcomes visible to a diverse set of communities, including academia and administration.

I am also grateful to my other committee members for their constructive and invaluable feedback and encouragement. William W. Cohen's courses on machine learning applied to text data were an invaluable learning experience for me as they equipped me with a rich set of skills and knowledge that was essential for my dissertation and beyond. Carolyn P. Rosé's door was always open for me, and our in-depth discussions about my research were crucial for me to scrutinize the validity of my methodologies and findings from different angles. Jeffrey Johnson has been an inspiring and motivating source of knowledge about social networks and their implications.

I owe a great deal of thanks to other faculty, staff and students from Carnegie Mellon. The discussions and collaborations with people in the CASOS lab and the COS program, especially Jim Herbsleb, Terrill Frantz, George Davis, Peter Landwehr and Frank Kunkel, were instrumental. I highly appreciate the intellectual interactions with David Krackhardt, Laura Dabbish and Keith Hunter from the Heinz College. CMU's administrative and support staff greatly contributed to my doctoral work. I thank Connie Herold, who helped with the logistics of the COS program, Monika DeReno, Janice Kusmierek, and Rochelle Economou, who worked at the CASOS lab, Catherine Copetas, Ed Walter, and the CS help desk. Anita Sarma and James Howieson from ISR provided valuable career advice.

Special thanks belong to the Women@SCS group at CMU who were an integrative force for me. I am especially grateful to Carol Frieze for enabling and supporting many of my computing outreach activities, and to Kenny Joseph and Ed McFowland for running the Computational Thinking middle school program with me.

I have greatly benefitted from the advice and support from extraordinary individuals from outside of CMU. The intellectual discussions with Harald Katzmair from FAS Research have broadened and enriched my understanding of society and sharpened my curiosity about human behavior. Corinne Coen from Case Western has been a mentor and friend.

My friends from Pittsburgh helped me to juggle my work-life balance. I am particularly grateful to Laurie Jones, Virginia Bedford, Kitrina Kearfott and Eric Malloy.

Last but not least, I am grateful to my personal network for their caring and loving support. In particular, I would like to thank my parents, Annelie and Rolf Diesner, for their moral support and good humor throughout my education. Finally, I owe a huge debt of gratitude to my partner Jan Reinhardt, who stood by my side with an infinite amount of encouragement, compassion and kindness.

Contents

Tables

Figures

1 Introduction and Overview

1.1 Thesis Statement

This thesis is motivated by the need for scalable, robust and reliable methods and technologies that support the construction of network data from natural language text data, and the usage of the extracted data for answering substantive and graph-theoretical questions about socio-technical networks. The findings and technology resulting from this thesis improve the applicability of language technologies for generating network data based on text data; thereby advancing the intersection of network analysis and text analysis. This thesis contributes to the actionable meaning of network data by providing methods that leverage theories from the social sciences to construct and analyze network data, and to combine text data and network data for analysis.

1.2 Network Analysis

Socio-technical networks represent interactions between social agents, infrastructures and information (Carley, 2002a). These networks are ubiquitous and impact society on many dimensions (Newman, 2010). Realizing the relevance of networks as a form of organizations and organizing, people from various backgrounds and domains, including academia, administration and business, have been asking questions such as:

- How can we efficiently and effectively collect, manage and analyze data about socio-technical networks in order to capture and understand the relevant properties and behavior of networks?
- What are the underlying forces that drive the evolution and dynamics of networks?
- What are the implications of certain network characteristics for practical purposes, such as building and managing teams and organizations, designing and adapting policies, disseminating information, and fostering innovation?
- How reliable are network analysis results?

In the field of network analysis, people have developed methods, metrics and theories that help to address these questions (Brandes & Erlebach, 2005; Freeman, 2004; Leinhardt, 1977). More specifically, *Social Network Analysis* (SNA) provides “a framework for testing theories about structured social relationships” (Wasserman & Faust, 1994, p. 17). Originally, SNA has been advanced by social scientists who used this approach to gain a rich and thorough understanding of small groups in a retrospective fashion (J. Mitchell, 1969; Newcomb, 1961; B. Ryan & Gross, 1943; Sampson, 1968). Therefore, the original network analytical measures were defined for connections between social agents, i.e. people and groups (Bonacich, 1987; Freeman, 1979; Wasserman & Faust, 1994). The scope of network analysis as a research method and of social networks as an object of study has been continuously broadened and adopted across disciplines. Consequently, a large body of new models, theories, methodological advances and applications has been developed (see for example Carrington, Scott, & Wasserman, 2005).

Network analysis is sometimes also referred to as *Network Science*, which is defined as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena" (National_Research_Council, 2005, p. 28). In network science, synthetic as well as empirical data are often used to study the quantitative properties, structure and dynamics of relational data (see for example Barabási & Albert, 1999; Erdős & Rényi, 1959; Simon, 1955; Watts & Strogatz, 1998). Network scientists have developed a wide range of efficient and scalable computational solutions for collecting, managing, and analyzing relational data (see for example Newman, Barabasi, & Watts, 2006). I herein refer to both SNA and Network Science, which are different labels for the same field, namely the study of relational or network data, as *network analysis*.

Based on the concept of *socio-technical systems* (Emery & Trist, 1960), the web of interactions within complex societal systems and their infrastructures is referred to as *socio-technical networks*. Most socio-technical networks exhibit characteristics of *complex systems*: they are in flux, vary in size, and feature a multitude of interactions and interdependencies between variables. This complexity can lead to radical changes in the system's behavior (Kauffman, 1995). The concept of socio-technical networks includes virtual and online networks.

In summary, network analysis has been adopted by researchers and practitioners as a general utility method – much like statistics – in a variety of fields, including business and economics (Burt & Janicik, 1996; Saaty, 2005), public policy (Krackhardt, 1990), social science and anthropology (Carley, 2002a; Johnson, Boster, & Palinkas, 2003), and computing (Balasubramanyan, Lin, & Cohen, 2010; Leskovec, Kleinberg, & Faloutsos, 2007). Furthermore, networks, especially social networks, have become a popular object of study (Newman et al., 2006).

1.2.1 Network Metrics

Core network metrics were developed with respect to social networks, i.e. people to people connections. In general, network metrics are defined on the node level, graph level, or the level of aggregates of nodes, including dyads, triads and clusters. The set of core metrics includes:

Node level: Centrality, which measures the prominence of a node with respect to the number of its direct connections (degree centrality), its distance to other nodes in the network (closeness centrality), how often it is positioned on the shortest path between any pair of nodes (betweenness centrality), and how close it is to other prominent nodes (eigenvector centrality) (Bonacich, 1987; Freeman, 1979).

Local level, aggregates of nodes:

- Dyads, i.e. two nodes and their connections: Structural and regular equivalence, i.e. which nodes show the same "network fingerprint" or are linked to the exact same set of other nodes, respectively (Everett & Borgatti, 1994; D. R. White & Reitz, 1983).
- Triads and higher order aggregates: The number of triangles, simmelian ties (directed edges in triangles), and cliques (maximally connected subgraphs) that an agent is involved in or that are present in a network (Krackhardt, 1998; Wasserman & Faust, 1994).

Graph level:

- The abovementioned centrality metrics are also defined on the graph level, where they are based on the respective centrality score nodes in the network, among other properties (Wasserman & Faust, 1994).
- Density, which measures the ratio of realized links to possible links (Wasserman & Faust, 1994).

A more complete definition of these metrics and all other metrics used in this thesis is provided in Table 154 in the Appendix. That Table also specifies which metrics are appropriate for what kind of network data. While the abovementioned metrics can be used for networks that involve any node class, network metrics have also been developed and defined for specific node classes (Carley, 2002b; Krackhardt & Carley, 1998). For example, the " knowledge load" metric measures the average number of nodes from the knowledge class that an agent is linked to (Carley, 2002b).

1.3 Network Data

Data on socio-technical networks can be collected through a variety of methods; most of which can be categorized as surveys (Krackhardt, 1987; B. Ryan & Gross, 1943), questionnaires (Newcomb, 1961), (participating) observations (J. Mitchell, 1969; Sampson, 1968), experiments (Milgram, 1967), and simulations (Carley, 1991). These methods can be conducted in a manual or computer-assisted fashion (Bernard et al., 1990).

Traditionally, researchers have used methods that required first-hand experience or direct interactions with network participants, such as (computer-assisted) personal and telephone interviews (Newcomb, 1961) and pile sorting (Boster, Johnson, & Weller, 1987). Even though these methods are expensive in term of costs for time and trained personnel, they have been widely used across various disciplines, including sociology (Bernard et al., 1990), anthropology (Bernard et al., 1990; Johnson et al., 2003; J. Mitchell, 1969), linguistics (J. Milroy & Milroy, 1985), political science (Hämmerli, Gattiker, & Weyermann, 2006), public policy and organization science (Krackhardt, 1990), and business (Galaskiewicz & Burt, 1991).

Over the last decade, network data collection methods have been adopted for online settings. Lately, harvesting the (participatory) web has become a widely used strategy for gathering network data (Parastatidis, Viegas, & Hey, 2009). Popular data sources include websites (Gloor et al., 2009), social networking sites such as Facebook and Twitter (Lampe, Ellison, & Steinfield, 2007), and other platforms for social interaction, such as blogs (Adar & Adamic, 2005), chats (Paolillo, 1999), and virtual worlds including online games (Bainbridge, 2007; Keegan, Ahmed, Williams, Srivastava, & Contractor, 2010).

1.3.1 Text Data as a Source for Network Data

The functioning and evolution of socio-technical networks involves the frequent production, processing and flow of information. This information often occurs in the form of natural language text data, and can originate from within or outside the socio-technical network of interest. It has been long recognized that such text data can serve as a single or complementary source of information about networks (Burt & Lin, 1977; Carley & Palmquist, 1991; Glaser & Strauss, 1967; Janas & Schwind, 1979). The availability of this type of data has stimulated a long tradition in linking text analysis and network analysis; with most of the prior research falling into one or more of the following categories:

- Analyzing semantic networks (for a review see Van Atteveldt, 2008).
- Defining network metrics for assessing relational data distilled from texts (Carley, 1997b).
- Developing methods, data structures and technologies for extracting relational data from texts (for reviews see Diesner & Carley, 2010c; Mihalcea & Radev, 2011).

Examples for types of the text data that have been used for network analysis include news wire data (Johnson & Krempel, 2004; Van Atteveldt, 2008), legal documents (Baker & Faulkner, 1993; Feldman & Seibel, 2006), transcripts of interviews and meetings (Carley, 1988; Dabbish, Towne, Diesner, & Herbsleb, 2011; Sageman, 2004), interpersonal communication such as traditional and electronic mail (Diesner, Frantz, & Carley, 2005; Fitzmaurice, 2000), and archival and historic data (Burt & Lin, 1977). More recently, text data that were generated as byproducts of (computer-supported) collaboration processes have become a popular source for collecting network data. Examples include descriptions of work processes (Corman, Kuhn, McPhee, & Dooley, 2002; Danowski & Edison-Swift, 1985), job training scenarios (Weil et al., 2008), e-learning environments (Haythornthwaite, 2001), team meetings (Dabbish et al., 2011), software development initiatives (Cataldo & Herbsleb, 2008), wikis (Chang, Boyd-Graber, & Blei, 2009), and virtual worlds such as online games (Landwehr, Diesner, & Carley, 2009).

In general, people mainly have been extracting three types of information from text data: First, one-mode networks, where nodes represent salient information from a corpus and are typically of the same node type. The resulting networks are often called concept networks (for a review see Diesner & Carley, 2010a). Concepts are considered as abstract representations of the information that people conceive in their minds (Sowa, 1984). Sometimes, concept networks are also called semantic networks, even though semantic networks are defined more strictly (Allen & Frisch, 1982; Sowa, 1992; Woods, 1975). Concept networks have been used to answer questions like: What are the key concepts in corpus? How do memes and ideas emerge, spread and vanish in society and on the internet? How do such diffusion processes happen over time? (Corman et al., 2002; Doerfel & Barnett, 1999; Gloor et al., 2009; Griffiths, Steyvers, & Tenenbaum, 2007; Leskovec, Backstrom, & Kleinberg, 2009)

Second, networks in which nodes represent entities of socio-technical systems, such as agents, locations and resources (Barthelemy, Chow, & Eliassi-Rad, 2005; Diesner & Carley, 2011b). Such (multi-mode) networks are also referred to as meta-networks (Carley, 2002a), and have been used to answer questions like: Who is talking to whom about what? Who are the key players in an organization? How does an agents' prominence differ depending on their access to resources and knowledge? What benefits and risks result from an observed network structure for the network and its wider context? (Carley, Diesner, Reminga, & Tsvetovat, 2007; Hämmerli et al., 2006; Van Atteveldt, 2008)

Third, texts or documents can also be considered as a node class themselves. These nodes can be linked to the social agents who have authored or cited a text or who are referenced in the data (Hummon & Doreian, 1989; C. Roth, 2006). Attributes of text data, e.g. meta-data such as index terms, can serve as additional nodes or node attributes (Pfeffer & Carley, under review). Networks in which text are considered as nodes have been used to ask questions like: Who has what impact on the advance of an idea or a discipline? How does co-publishing within versus across organizations relate to the acquisition of research funding? (Small, 1973; Wagner & Leydesdorff, 2005)

Overall, network analysis has been used on unstructured, semi-structured and structured natural language text data. Unstructured means that only plain text bodies are available. Semi-structured means that chunks or tokens in the data are annotated with additional information, such as turns between speakers. Structured means that the text bodies are annotated such that they allow for filling templates that have a predefined structure, such as tables and databases, or that the annotations adhere to a predefined taxonomy or ontology.

1.4 Opportunities and Challenges of Bringing Together Text Analysis and Network Analysis

Historically, hand coding has been a dominant way for coding texts as networks (Bernard & Ryan, 1998; Glaser & Strauss, 1967; Novak & Cañas, 2008). Due to technical advances, the storage and retrieval of text data with information about networks has become fast, cheap and easy (Shapiro, 1971; Trigg & Weiser, 1986). Modern information and communication technologies, such as the internet, cell phones, and social networking services, have further expedited and facilitated the production, distribution and collection of a) network data and b) text data pertaining to networks (Eagle & Pentland, 2006; Parastatidis et al., 2009). Since hand coding does not scale up the amount of text data available for analysis, there is a broad need among researchers and practitioners for theories, methods, metrics, and tools that support efficient knowledge discovery and reasoning about network data extracted from text data (Carley, 2002a; Schrodt, 2001; Shen, Ma, & Eliassi-Rad, 2006). At a minimum, end users are interested in text mining solutions that help them to gain a first pass understanding of the properties and dynamics of socio-technical networks (Bond, Bond, Oh, Jenkins, & Taylor, 2003; McCallum, 2005; Parastatidis et al., 2009). Such data are often used as a starting point for further analysis such as close readings. In addition to this purpose, people have been using network data extracted from texts for the following purposes:

- Populating relational databases, which can be used for information search and retrieval purposes (Brin, 1999; Cafarella, Banko, & Etzioni, 2006; Fellbaum, 1998; Gerner, Schrodt, Francisco, & Weddle, 1994; King & Lowe, 2003).
- Input to further computations, such as simulations of socio-technical systems and machine learning procedures (Carley et al., 2007; Pearl, 1988).
- Generating network visualizations, e.g. for engaging people in communication about complex systems and conflicts (Hämmerli et al., 2006; Hartley & Barnden, 1997; Johnson & Krempel, 2004; Shen et al., 2006).
- Iterative testing and development of theories about socio-technical systems (Glaser & Strauss, 1967; J. Milroy & Milroy, 1985).

- Monitoring and improving organizational and collaborative processes (Corman et al., 2002; Dabbish et al., 2011; Weil et al., 2008).
- Assessment of conflict escalations and early warning systems for crises, as well as a data source for analyzing crises (Bond et al., 2003; Hämmerli et al., 2006; Zagorecki, Ko, & Comfort).

Even though the combination of text analysis and network analysis has led to advances in research and practical applications in either field, it also involves unique challenges. Some of these challenges are addressed in this thesis:

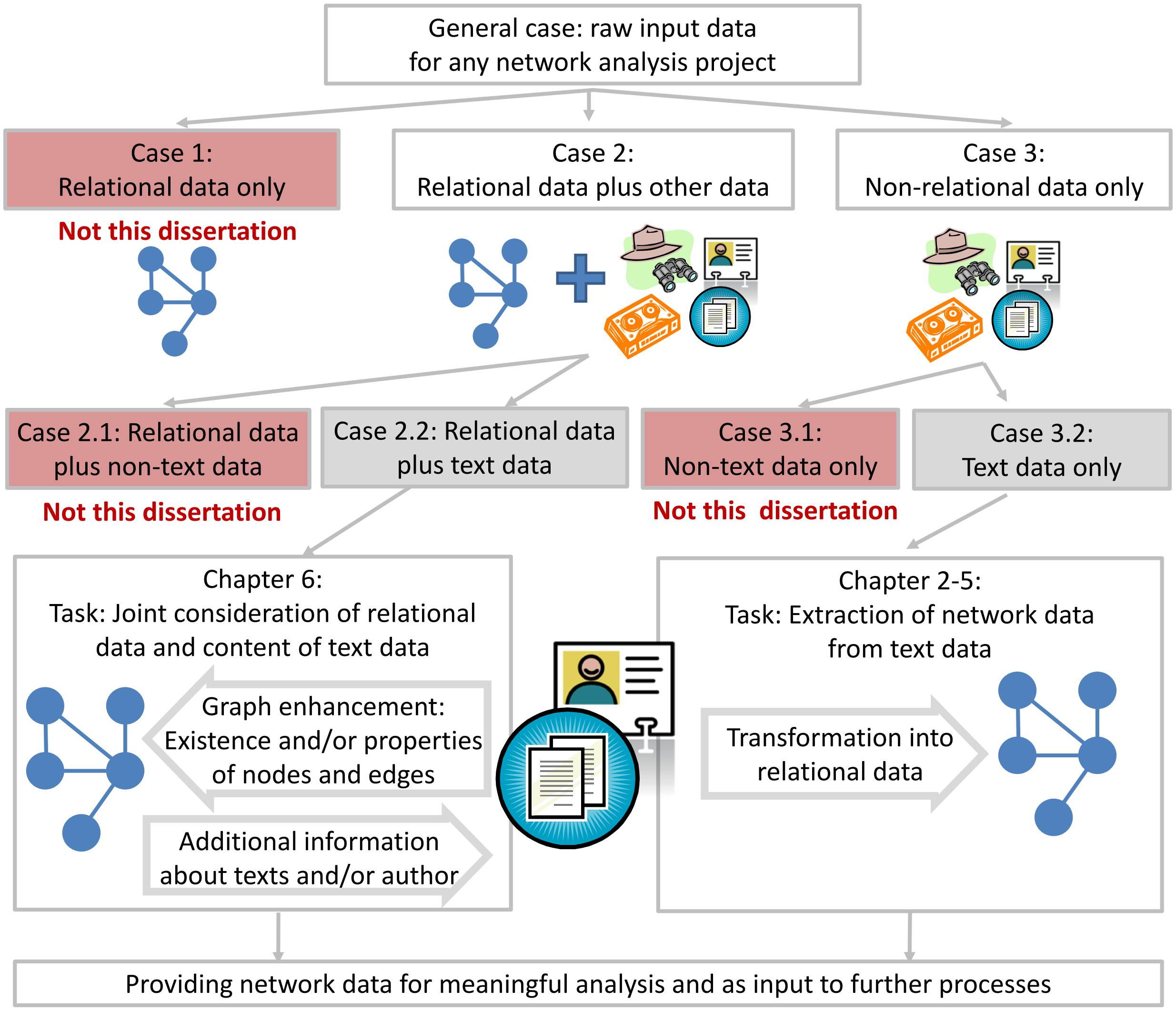
- The efficient and reliable extraction of nodes and links from text data (Corman et al., 2002; McCallum, 2005). This issue mainly applies to unstructured text data.
- The lack of sufficient amounts of (reliable) ground truth that can be used for validating network data extracted from texts. This challenge applies to unstructured, semi-structured and structured text data.
- The fusion of unstructured and structured information from text data.

Besides these challenges, there are many others, which are beyond the scope of this thesis. Examples include biases in texts, emotions and sentiments expressed by members of social networks (Shanahan, Qu, & Wiebe, 2006), and adapting existing methods and tools to new domains and genres (Gupta & Sarawagi, 2009), such as social media data and email data (McCallum, Wang, & Mohanty, 2007).

1.5 Organization of Thesis

The chapters in this thesis are organized by the different types of availability of text data for network analysis, and the structuring of these text data; going from the availability of unstructured text data only (chapters 2 - 4) to (semi-)structured text data plus other sources for network data (chapters 5, 6). These different options are depicted in Figure 1 and described below. Table 2 summarizes which type of structuring is addressed in which chapter, and which types of structuring the respective findings apply to.

Figure 1: Organization of thesis*


General case: raw input data
for any network analysis project
Case 1:
Relational data only
Not this dissertation
Case 2:
Relational data plus other data
Case 3:
Non-relational data only
Case 2.1: Relational data
plus non-text data
Not this dissertation
Case 2.2: Relational data
plus text data
Case 3.1:
Non-text data only
Not this dissertation
Case 3.2:
Text data only
Chapter 6:
Task: Joint consideration of relational
data and content of text data
Graph enhancement:
Existence and/or properties
of nodes and edges
Additional information
about texts and/or author
Chapter 2-5:
Task: Extraction of network data
from text data
Transformation into
relational data
Providing network data for meaningful analysis and as input to further processes

* Gray fields: situations addressed in thesis. Red fields: situations not considered in thesis

Availability of text data only (Figure 1, case 3.2): The structure and behavior of networks can be explicitly or implicitly encoded in text data. Sometimes, texts are the only source of information about a network. Most of these cases fall into one or more of the following categories, which are not exclusive:

- Networks that are inaccessible or unobservable for researchers:
    - Covert networks, e.g. illegal business coalitions (Baker & Faulkner, 1993) and adversarial groups of sub-state and non-state actors (Krebs, 2002; Sageman, 2004).
    - Networks that do not exist anymore, e.g. former regimes (Seibel & Raab, 2003) and bankrupt companies (Diesner et al., 2005).

- Virtual networks that are not based on an underlying real-world network, or that are nothing more than the data traces produced in these networks, such as blogs (Adar & Adamic, 2005). We refer to such networks as WYSIWII (What-You-See-Is-What-It-Is) (Diesner & Carley, 2009b).
- Very large networks, where conducting surveys within appropriate network boundaries would be prohibitively expensive (Burt & Lin, 1977), e.g. geopolitical networks.
- Groups that do not produce large amounts of readily available interaction data, e.g. ethnic groups (J. Mitchell, 1969), or interactions in offline, not computer-supported settings.
- Semantic networks and network representations of mental models, i.e. structured representations of information that people conceive in their minds (Klimoski & Mohammed, 1994; Rouse & Morris, 1986).

In these cases, network data can be extracted from text data. From an NLP point of view, this is an Information Extraction (IE) task referred to as Relation Extraction (REX) (McCallum, 2005). REX is particularly valuable when text data are the only source of information about a network. However, the network data resulting from REX are hard to verify when (reliable) ground truth data are missing (Klerks, 2001). This is often the case for covert and large-scale networks, for example. This limitation is even more severe if we consider the fact that the computational and interdependent steps needed for highly accurate REX solutions impact the structure and properties of the distilled network data. These impacts are insufficiently understood (Corman et al., 2002). I start to bridge this knowledge gap in chapter 2, where I investigate the amount and boundaries of variation in network structure that is due to engineering decisions made when building relation extraction tools and end-users decisions made when applying these tools.

In the social sciences, people have developed theoretically grounded and empirically tested models of socio-technical networks. These models can be used as ontologies for defining the entity classes that are relevant for Information Extraction and REX (Barthelemy et al., 2005; Van Atteveldt, 2008). One of these models is the meta-matrix model, which contains entity classes including and beyond the set of classes typically considered for REX, i.e. people, organizations and locations (Carley, 2002a; Krackhardt & Carley, 1998). However, there is a lack of:

1. Technologies that facilitate the efficient extraction of network data that adhere to the meta-matrix model.
2. Evaluations of the performance of such extraction technologies in practical applications settings beyond experimental studies that serve the formal model validation based on ground truth data.

The first need is addressed in chapter 3, where I develop and evaluate prediction models for entity extraction. These models distill instances of meta-matrix entity classes from unstructured text data. The retrieved entities can be used as nodes for constructing socio-technical networks. In chapter 4, I describe how the developed entity prediction models are integrated into an end-user software product as well as the operational implications of this process.

The second need is addressed in chapter 5, where I evaluate the performance of the aforementioned prediction models in different, practical application contexts. In that chapter, I also compare the resulting networks with respect to their structure and properties to networks generated with alternative methods from the same text data. The ultimate goal with this work is to provide network data that can be used to answer substantive and graph-theoretical questions about socio-technical networks. The comprehensive analyses needed to answer the first type of questions require additional empirical studies, which are beyond the scope of the thesis. The point with this chapter is rather is to illustrate the process of going from a research question to the collection and analysis of network data. I describe the methodological steps and choices involved in this process such that this information can serve others as a guideline for conducting empirical studies.

Joint availability of text data and network data (Figure 1, case 2.2): Sometimes, in addition to text data, further sources of information about a network are available, such as relational data or meta-data from which relational data can be constructed. Prominent examples for this situation include:

- Surveys that ask respondents not only for information about entities and relations (relational data) (see for example Krackhardt, 1987), but also for answers to questions that further describe the nature of the nodes and links (text data) (Palmquist, Carley, & Dale, 1997).
- Communication networks (who is talking to whom, relational data) about what (text data) (Monge & Contractor, 2003).
- Co-citation networks, where person *A* is linked to person *B* if *A* cited *B* (relational data) in a paper (text data) or patent (Hummon & Doreian, 1989; C. Roth & Cointet, 2010).
- Web science studies that combine data on the connectivity between URIs (relational data) with the content of the corresponding webpages (text data) (Adar & Adamic, 2005; Kleinberg, 2003).

Two approaches are commonly used for representing and analyzing both types of data: First, the text data and the relational data are analyzed separately from each other. Second, the text data are reduced to the fact, frequency or likelihood of the flow of information between nodes. This is typically done by representing the exchange of information as a link. While the second approach is efficient and acknowledges that information exchange has taken place, it does not consider the substance of text data. However, we know that without considering the content of text data, or by analyzing text data and other data about a network in a disjoint fashion, we are limited in our ability to understand the effects of language use in networks. This includes the transformative role that language can play in networks, and the interplay and co-evolution of information, and the structure and behavior of networks (Corman et al., 2002; Danowski, 1993). Approaches to considering the content of texts are built on the idea that "travelling through the network are fleets of social objects" (Danowski, 1993, p. 198), where these objects can be language, norms, practices, and other types of human behavior and social interactions (Bourdieu, 1991; Eckert, 2000). The lack of integration and joint analysis of text data and other types of data about networks is addressed in two places in this thesis: First, in chapter 5, where I show to what extent networks extracted from texts data agree in structure and key entities with networks built from meta-data or in collaboration with subject matter experts Second, in chapter 6, where I propose and demonstrate a methodology for jointly considering relational data and text data.

Finally, text data sources may also contain non-textual information that are not addressed herein, such as images, audio and video data (Figure 1, Case 2.1). These additional types of data might contain further information about networks. While I do not consider these alternative types of non-relational data herein, the methods for and insights from comparing and integrating text networks and networks from other sources might serve others as a starting point for bringing together different types of information about networks.

1.5.1 Datasets Used in Thesis

For the experimental work in chapters 2 and 3, I use external, validated, ground truth corpora. With this kind of data, I am able to measure the actual and precise impact of coding choices on network data, and to validate the prediction models in a reliable and controlled fashion. These datasets are introduced in chapter 2.

For the applied work in chapters 5 and 6, I use a corpus that we have previously collected (Enron) (Diesner et al., 2005), and two corpora that I have collected and prepared for this thesis (Sudan, Funding). The Enron data contain emails from employees in the Enron corporation (Diesner et al., 2005). The Sudan corpus consists of news wire articles about the Sudan, plus meta-data on these articles, such as publication dates and index terms. The Funding corpus comprises proposals of funded research projects, plus information about the people involved in these projects, and additional details about the projects, such as amount of funding awarded. These datasets are introduced in detail in chapter 5. Table 1 compares the datasets used herein along various characteristics. Even though the datasets are from different domains - namely industry, politics, and science - they share a few characteristic:

- All datasets contain natural language text data.
- All datasets contain some meta-data.
- All datasets contain time-stamped, long-term, over-time data.

Much of the recent work on combining text analysis and network analysis investigates the properties and benefits of interaction between humans via social media and computer supported collaborative work environments. In contrast to that, the datasets used herein represent networks that involve conflicts (Enron, Sudan) and competition (Funding). Prior research suggests that for such networks, the formation and cohesion of groups might also be driven by external pressures, such as scarce resources and the struggle for power, more so than by group-inherent characteristics, such as a shared identity and the desire to collaborate. These properties have shown to foster the development of strategic alliances (Fitzmaurice, 2000). For situations in which groups need to balance concealment and coordination, prior research has provided empirical evidence for how these networks differ from overt networks (Baker & Faulkner, 1993). However, this thesis is focused on methodological questions instead of substantive questions about the considered datasets and domains. Nonetheless, the technologies and methods developed and evaluated herein are tested on these datasets, such that the gained insights can be expected to generalize within the stated boundaries to other datasets from similar domains. This helps to complement knowledge about classic cooperation and collaboration networks, and addresses shortcomings with methodological issues for analyzing covert networks (Klerks, 2001; Skillicorn, 2008).

Table 1: Comparison of datasets

Dimension	**Sudan Corpus**	**Funding Corpus**	**Enron Corpus**
Domain	Geo-political : Politics, conflict, covert activities	Science: Innovation, collaboration, competition	Business: Innovation, politics, covert activities
Social network	Implicit in text bodies	Explicit in project descriptions	Explicit in emails headers
Semantic information/ network	Implicit in texts	Implicit in abstracts	Implicit in email bodies
Size	79,388 articles	55,972 proposals	52,866 emails
Time span	12 years	25 years	6 years
Original access to data	Public	Beginning: internal If funded: public	Internal

Intended audience	The public Analysts	Program managers Scientific community	Addressees
Style	Formal: journalistic	Formal: scientific	Formal and informal

Table 2: Types of text data and networks used in thesis*

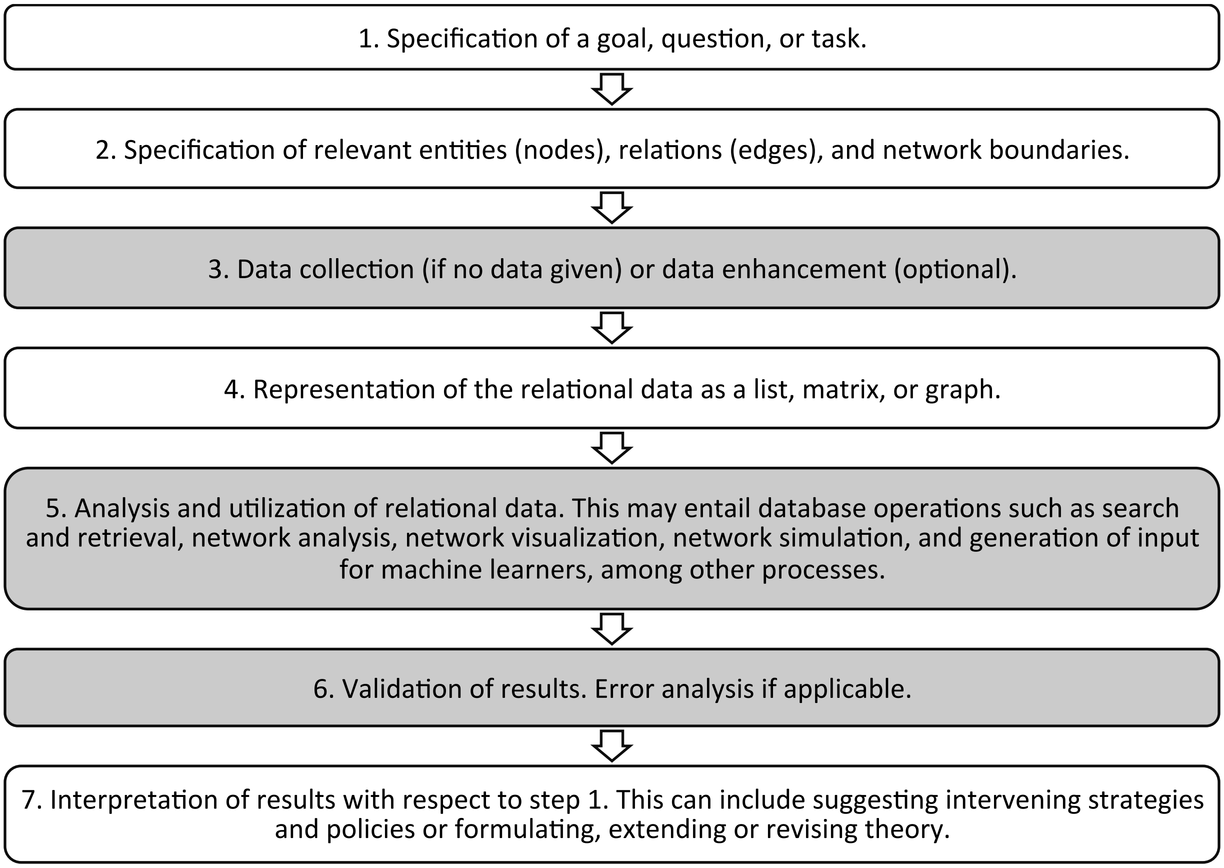
Chapter	Experiments and Analyses		Insights gained and technology built applicable to	
	Network modality	Type of structure of text data	Network modality	Type of structure of text data
2: Investigation of impact of coding choices on network structure and network analysis results.	One-mode networks (reference resolution project, windowing project). Multi-mode networks (windowing project).	Unstructured	One-mode networks and multi-mode networks.	Mainly unstructured data. Also applicable to structured data.
3. Entity Extraction for providing nodes for constructing socio-technical networks.	One-mode networks and multi-mode networks.			
5. Comparison of networks generated with various relation extraction techniques.		Unstructured: Sudan: news articles Funding: research proposal Enron: email bodies Structured: Sudan: meta-data Funding: meta-data Enron: email headers		
6. Method for combining content of text data with social network data.	One-mode networks of different modes (concept network, social network).			Unstructured data for which meta-data are also available.

* Using the definition of structured and unstructured data presented in this chapter, most data annotated for information extraction purposes falls under the category of structured data. However, the actual texts in such datasets are unstructured. Entries marked with a * in this table represented cases in which unstructured text data with annotations that bring some form of structure to the text are used.

1.6 The Network Analysis Process

The questions addressed in this thesis relate to certain steps of the overall network analysis process. Since network analysis has originated from various fields with cross-disciplinary influences, the methodology for conducting network analysis is less standardized than research methodologies that are more specific to a field. Synthesizing prior descriptions of the network analysis process (Knoke & Yang, 2008; Wasserman & Faust, 1994) suggests that this process comprises the seven steps shown in Figure 2. In this figure, the steps towards which this thesis makes a contribution are marked with gray backgrounds. Since these individual steps are highly interdependent, any individual step can be assumed to have recuperations on the following steps as well as on the overall outcome of a network analysis project.

Figure 2: Network analysis process and steps focused on in this thesis (gray)


1. Specification of a goal, question, or task.
2. Specification of relevant entities (nodes), relations (edges), and network boundaries.
3. Data collection (if no data given) or data enhancement (optional).
4. Representation of the relational data as a list, matrix, or graph.
5. Analysis and utilization of relational data. This may entail database operations such as search and retrieval, network analysis, network visualization, network simulation, and generation of input for machine learners, among other processes.
6. Validation of results. Error analysis if applicable.
7. Interpretation of results with respect to step 1. This can include suggesting intervening strategies and policies or formulating, extending or revising theory.

1.7 From Text Data to Network Data to Knowledge

The focus of this thesis is on the collection, analysis and validation of network data extracted from texts. I distinguish between network data and relational data. What is the difference, and why does it matter?

Relational data, also referred to as *graphs*, consist of vertices, also called nodes, and of edges, also called arcs, links, or connections. The edges connect the nodes. Additionally, nodes and edges can have weights, attributes, types, and probabilities, and links can furthermore have directions. Nodes can represent instances of one (one-mode) or more (multi-mode) types of entity classes, such as "agent" and "information". Edges can represent instances of one (uniplex) or more (multi-plex) types of relationships, such as “collaboration” or “trade” (Carley, 2002a; Wasserman & Faust, 1994, p. 79). *Social networks*, for example, involve only entities of the type “agent”.

Network data consists of relational data plus additional data that help to contextualize and interpret relational data (Alderson, 2008). Thus, relational data are an indispensable subset of network data, but are insufficient for revealing comprehensive stories about socio-technical networks (Corman et al., 2002).

It has been previously argued that in order to allow for meaningful analysis of socio-technical networks and for answering substantive questions about such networks, linked data need to be transformed into information, and information into knowledge (Parastatidis et al., 2009). Translating this argument into network terms means to go from relational data to network data, and from network data to knowledge. Transforming relational data into network data requires the enhancement of relational data with additional data (Alderson, 2008). This is typically achieved by bringing together various types or sources of information about a network. This theoretical argument has been put into action by applying one or more of the following strategies:

- Including attributes that describe relevant characteristics of nodes and/or edges (Sampson, 1968).
- Considering different views of a network (Krackhardt, 1987).
- Enhancing relational data with additional data that help to fix the context of the relational data.

Additional data about networks are often referred to as meta-data. Widely adopted types of meta-data are temporal and spatial information, such as timestamps of events or the geophysical position of nodes (Eagle & Pentland, 2006; Snijders, 2001). Another type of additional data are natural language text data (Carley & Palmquist, 1991; Danowski, 1993). This thesis focuses on the latter option, i.e. using text data to construct and enrich relational data and network data. While texts generated by humans can be considered as a type of behavioral data, meta-data can be generated by humans or automatically, e.g. in the case of key words for documents. This thesis is focused on methods for utilizing human-generated text data pertaining to socio-technical networks, including meta-data.

Going from networks to knowledge means to perform analyses such that substantive questions about networks can be answered. In general, this requires the usage of methods and the computation of metrics that are appropriate for some given network data. Sometimes, using generic matrix operations or calculating metrics that are defined independently of the type of nodes or edges is most appropriate and sufficient. This often applies to research problems in network science. In other cases, methods and metrics are needed that take the types or other characteristics of nodes and edges into account (Carley, 2002a; Krackhardt & Carley, 1998). This can apply to the analysis of multi-mode or multi-plex networks, for instance (Cataldo, Herbsleb, & Carley, 2008; Krackhardt & Carley, 1998). When the latter approach is more appropriate, there are several models and metrics available that are based on theories about the system that the network data represent. I herein follow this route by using a theoretically grounded model of socio-technical networks to inform the selection of entity types to extract network data from text data.

In summary, going from relational data to network data to knowledge helps to make the substance or meaning of network data practically useful and actionable. Here, "practically useful" and actionable "actionable" means extractable, explicitly representable, and useful for answering substantive and graph-theoretical questions about socio-technical networks. Sometimes, this process is used to develop strategies for taking further actions, such as designing and implementing policies and interventions. The concept of *actionable meaning* as used in this thesis is closely related to *semantic computing*, which refers to "computing with (machine processable) descriptions of content and intentions" (Parastatidis et al., 2009). The difference between semantic computing and making the substance or meaning of network data actionable is that the approach followed herein does not necessarily imply the consideration of intensions, but focuses on contributing to the potential practical usefulness of network data.

1.8 Summary of Contributions

The study of the impact of coding choices on network data and analysis results (chapter 2) and the implications of these findings for practical work (chapter 4.1) can help people to become better informed users of relation extraction methods and technologies, to gain greater control over these multi-step analysis procedures, and to draw reasonable conclusions from network analysis results. The findings from chapter 2 emphasize that it is crucial to know the amount and nature of the impact and interaction effects of routines involved in relation extraction on network data. This work together with the testing of the prediction quality of an entity extractor (chapter 3) in different applications settings (chapter 5) complements traditional methods for assessing the accuracy of relation extraction methods.

In chapter 4, the transition from experimental results to a) the impact of coding choices on network data and b) the accuracy of the entity extractor in real-world applications is described. This work increases the practical usefulness and interpretability of network analysis results. Also, the challenges identified for converting trained prediction models into ready to use software, and the developed solutions to these challenges can provide others with guidance for this kind of design and engineering problem.

Based on the comparison of network data generated with different methods from the same corpora (chapter 5), the differences and commonalities in network structure and analysis results are identified. Moreover, I show which findings generalize across domains and writing styles, and which ones are more domain-specific. This knowledge is relevant in the context of networks for which insufficient or unreliable ground truth data are available, because in these situations, it is crucial to know how the views on a network differ depending on the employed relation extraction method. This work has also shown that generating thesauri by using the entity extractor built in chapters 3 and 4 greatly reduces the time costs for constructing thesauri in a manual or semi-automated fashion. However, based on the findings from the qualitative assessment of the auto-generated thesauri, it does not seem recommendable to use these thesauri without further verification and refinements. The strategies and tools for post-processing the auto-generated thesauri that I describe and develop in chapters 4 and 5 might help others with this process. Moreover, my results show that working through this refinement process increases the similarity between networks generated by using the auto-generated thesauri and networks generated with alternative methods.

In chapter 6, an advancement to the methodological approach of enhancing social network data with content nodes extracted from text bodies is developed, operationalized and tested in practical application scenarios. The proposed approach considers the substance of text data and helps to integrate different aspects that drive the properties and dynamics of networks. I conclude that extracting content nodes from groups of *structurally equivalent* agents is an appropriate strategy for enabling the *comparison* of the information that these agents produce, perceive or disseminate, while extracting content nodes from groups of *structurally coherent* agents is an appropriate strategy for enabling the *enhancement* of social network data with content nodes. The results from putting the latter approach to the test include a comparison of the outcome of topic modeling to the results from alternative information extraction methods, including entity extraction based on supervised learning. My findings show that performing key player analysis on text-based networks retrieves only a small portion of entities that would not be found with topic modeling, and that entities from meta-data knowledge networks might serve as suitable labels for unlabeled topics. Also, these comparisons further complement the findings from previous chapters about the differences and commonalities between various methods for constructing network data from text corpora.

In summary, by bringing together text data and relational data, this thesis makes substantial advances at the nexus of text analysis and network analysis. Using text data for network analysis is further a valuable strategy for contextualizing and interpreting graphs, and transforming linked data into useable information and knowledge (Parastatidis, et al., 2009).

2 Impact of Methodological Choices for Relation Extraction on Network Data and Network Analysis Results[1]

When network data are needed and text data are available as a source of information, network data can be extracted from text data. In computer science, this task is referred to as Relation Extraction (REX). Relational data extracted from text data may represent the nodes and edges in the network of interest accurately or not. This chapter makes a contribution by complementing traditional REX evaluation metrics with novel experimental results that help to answer the following questions:

- How do methodological choices for REX impact the extracted network data and respective analysis results?
- How do errors made during text data preprocessing, node identification and link identification propagate through REX tool chains?
- How much do improvements in REX accuracy rates help to diminish these effects?

Two particular coding choices are considered in this chapter: first, reference resolution, which identifies the entity or entities that a referring expression such a pronoun or co-reference refers to (sections 2.4, 2.7.1). Second, windowing, which is proximity-based method for linking extracted nodes (sections 2.5, 2.7.2). The results from this chapter inform us on how much the coding choices that one needs to make when extracting network data from text data as well as typical error for these procedures impact common network metrics and the identification of key players.

2.1 Introduction to Relation Extraction from Text Data

Methods for going from texts to networks have been developed in different fields, mainly Artificial Intelligence (AI) (Sowa, 1992), Natural Language Processing (NLP) and Computational Linguistics (CL) (Mihalcea & Radev, 2011), social science (Carley, 1993; Glaser & Strauss, 1967) and political science (Gerner et al., 1994). Even though these methods differ in their terminology, underlying theories and assumptions, degree of automation, evaluation strategies, and typical application areas, they overlap in that they exploit one or more of the following types of information:

- Lexical and morphological information, i.e. words and their structure (e.g. Woods, 1975).
- Syntax, i.e. the relationship between words (e.g. Janas & Schwind, 1979).

[1] In this chapter, portions are reprinted, with permission, from: Diesner, J., & Carley, K. M. (2009). He says, she says. Pat says, Ttricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis. Proceedings of IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA), Ottawa, Canada, © IEEE.

- Semantics, i.e. the meaning of words and language (e.g. Fillmore, 1968; Ogden & Richards, 1923).
- Pragmatics, i.e. the social use of language (e.g. Hovy, 1990).
- Logical (e.g. Shapiro, 1971) and statistical (e.g. Pearl, 1988) information.

These types of information are explicitly or implicitly contained in text data or can be inferred from it. Section 3.2 provides a problem-oriented review of the families of methods for going from texts to networks. For a more comprehensive review, see also Diesner and Carley (2010c). Currently, the most accurate, efficient and scalable REX methods combine NLP and CL techniques, and involve routines from statistics and machine learning (McCallum, 2005; Van Atteveldt, 2008). At a minimum, REX involves three steps, which are typically performed in the following order:

1. Data pre-processing: this includes subroutines such as chunking (partitioning texts into semantic units, typically sentences), reference resolution and word sense disambiguation.
2. Node identification, and if needed node classification: the generalized version of this task has been studied in NLP and Information Extraction (IE) under the label of Named Entity Recognition (NER) (Bikel, Schwartz, & Weischedel, 1999), and also in political science, where it is called event data coding (Schrodt, Yilmaz, Gerner, & Hermick, 2008). A more detailed introduction to this and the next step is provided in section 3.2.
3. Edge identification, and if needed edge classification: in this step, the identified nodes are linked into edges (Miller, Fox, Ramshaw, & Weischedel, 2000; Zelenko, Aone, & Richardella, 2003).

Tremendous progress in the automation and performance of REX has been achieved over the last decade (see for example Brin, 1999; Bunescu, 2007; Etzioni et al., 2004; McCallum, Wang, & Mohanty, 2007; Zelenko et al., 2003). These advances are mainly due to two reasons: First, they were facilitated by REX competitions that were initiated and funded by US-American governmental agencies, such as the Message Understanding Conference (MUC) (Chinchor, 2001), the Automatic Content Extraction Program (ACE) (Walker, Strassel, Medero, & Maeda, 2006), and the Translingual Information Detection, Extraction and Summarization Program (TIDES) (A. Mitchell et al., 2004). These competitions involve the provision of benchmark datasets and the development of REX evaluation metrics. Second, advances in REX have been attributed to progress with statistical and machine learning techniques, which have been developed and adopted by NLP researchers (Mihalcea & Radev, 2011).

2.2 Evaluation of Relation Extraction: Problem Statement

In Information Extraction, accuracy is typically measured as the percentage of correctly identified and classified items, in this case nodes and edges. Two methods are available for determining the accuracy of the retrieved data:

First, the "gold standard test", which compares distilled network data against ground truth data that has been previously annotated by trained human experts with entities and/or relations. The manual or computer-supported generation of correct and reliable ground truth data is expensive. For the domain of political event data coding, for example, it has been shown that humans trained for this task can identify and mark up about five to ten relations or events per hour, or up to 40 relations per day (Schrodt, 2001; Schrodt et al., 2008). Fortunately, various annotated datasets for IE tasks, including NER and REX, have been generated for nationally funded initiatives and made available through the Linguistic Data Consortium (LDC). An overview of these datasets is provided in Table 5. However, the non-trivial task of annotating data for REX has led to compromises: First, most standard REX datasets denote relations mainly on the sentence level (Bond et al., 2003). One explanation for this effect could be that the reliable identification, disambiguation and annotation of entities and relations within and across multiple sentences, paragraphs, documents or even corpora might be cognitively too complex for humans to perform (Corman et al., 2002). Second, the number of entity classes and even more so of node classes to be considered for REX is often kept fairly small: typically, solutions are developed that are constrained to locating and classifying entities that represent people, organizations and locations that are referred to by a name. For edges, most solutions are developed for relationships that are defined over the named node types, and sometimes also classify these relations according to a (predefined) ontology. As a result, the workflow in many of these systems is such that entities are identified first, and edges second. In an attempt to challenge this standard procedure, Roth and Yih (2002) showed that knowing the class label of entities helps to label relations, but not vice versa. These results confirm the traditional sequence of steps taken in REX.

As an alternative to the gold standard test, REX outputs can be assessed by subject matter experts (SMEs). The SMEs examine how closely the extracted data resemble the actual network of interest (King & Lowe, 2003). However, for real-world applications, the obtained network data are often too voluminous and too complex to be vetted by humans for accuracy. To make things worse, in some cases, neither any ground truth data nor SMEs might be available for validating results, e.g. when performing REX on historical data (Bearman & Stovel, 2000).

In NLP, accuracy is typically measured in terms of recall, precision, and a weighted average of these two metrics. The formulas for these metrics are given below. Recall measures coverage, i.e. what percentage of entities or links from the ground truth data have been retrieved. Precision measures accuracy, i.e. what percentage of the retrieved items, which can include false positives, are correct ones, i.e. occur in the ground truth data. Since recall and precision are typically inversely related, the harmonic mean of both values is also computed, which is called the F-measure.

Equation 1

$$Recall = \frac{number\ of\ correctly\ classified\ entities\ retrieved}{number\ of\ entities\ in\ ground\ truth}$$

Equation 2

$$Precision = \frac{number\ of\ correctly\ classified\ entities\ retrieved}{number\ of\ entities\ retrieved}$$

Equation 3

$$F = \frac{Recall * Precision}{0.5(\ Recall + Precision)}$$

There is a conceptual difference between ground truth and accuracy. In this thesis, I am working with two types of ground truth data: First, annotated corpora from external providers such as the Linguistic Data Consortium. For these data, intercoder-reliability, which is an approximation of the accuracy or quality of the data, has already been measured. Second, coding material and network data provided by SMEs. The assumption here is that the SMEs are able to provide an accurate assessment or picture of a socio-technical system. In general, it can be difficult to ascertain the truth of some text data, for example because this truth can only be elicited from the authors or is implicitly represented in the data. I make no claim that the ways in which I capture "ground truth" represent what the author intended or what others might extract with other compendiums. Rather, I claim that if other's use the same approaches to establishing ground truth they can compare the accuracy of their algorithms against those described herein. Accuracy is defined as the extent to which the data extracted from the text, i.e. the coded data, matches the ground truth data. That is, the extent to which the "concepts" extracted as nodes, the links, and the ontological classification generated by the algorithm are identical to those in the ground truth.

In summary, REX evaluation methods and metrics are tuned towards maximizing the accuracy of REX methods while avoiding overfitting to the training data. Here, accuracy means resemblance of the ground truth as identified by human experts; either expert data annotators or subject matter experts. As a consequence, research efforts in this area have been focused on improving existing REX methods or developing new ones, and reporting increases in accuracy over a baseline, established benchmark values, and alternative or competing systems. Typically, the research question asked with this type of work is, in a simplified form: How can we build a method, algorithm or technology that leads to the comparatively most accurate relation extraction results? I argue that while answers to this question advance the field of NLP, this common research question does not address two additional aspects of accuracy that are also crucial for understanding and improving the performance of solutions to REX. I elaborate on these two aspects in the following two subsections.

2.2.1 Impact of error propagation through interdependent subroutines on network data

First, the steps involved in REX, i.e. pre-processing and the identification (and classification) of nodes and edges, are not independent of each other, meaning that decisions made for one step can impact the results obtained from any subsequent step (Bernard & Ryan, 1998; Carley, 1993; C. W. Roberts, 1997b; D. Roth & Yih, 2002; Sarawagi, 2008). This type of complexity is further increased by the fact that modern REX techniques typically comprise multiple subroutines per step, and these subroutines can also exhibit interaction effects. The problem here is that even though the described interdependencies can lead to cascading errors and impact on intermediate and subsequent results, we do not have a good understanding of these effects, their impact on final results, and the robustness of REX methods towards these effects. One reason for this lack of knowledge is that this question has not yet been raised. This is troublesome because any error throughout the REX process can lead to inaccurate network data, erroneous analysis results, misleading interpretations, and unjustified further actions. Addressing this question becomes even more important when considering that the steps involved in REX are not flawless themselves: standard pre-processing techniques, such as part of speech tagging and reference resolution, have error rates of about 4% and 20% to 40%, respectively (Denis & Baldridge, 2007; Diesner & Carley, 2008b). For entity extraction, accuracy rates range from 80% to more than 90% (CoNLL-2003, 2003; MUC7, 2001). The edge identification stage will inherit both of these errors. Moreover, top performing relation extraction solutions have error rates of 30% up to 50% (Sarawagi, 2008). Yet another factor contributing to the limited understanding of interdependencies and error propagation in REX is that state of the art REX systems do not necessarily expose or provide detailed documentation on the employed subroutines. Therefore, the propagation of variation in results is not always transparent or comprehensible for the end user. Moreover, computational solutions for each of the three steps involved in REX are often developed independently from other steps. For example, link identification algorithms, methods and tools often assume that node identification has already happened (Chang, Boyd-Graber, & Blei, 2009). This separation of tasks inhibits the investigation of the end-to-end propagation of errors.

2.2.2 Impact of relation extraction subroutines on network analysis results

Second, the selection of specific methods and subroutines for REX impacts not only the accuracy of extracting entities and relations, but also the structure and properties of the retrieved relational data. However, the relationship between changes in the accuracy of REX and changes in network properties are also insufficiently investigated and understood, even though this gap in research has been previously pointed out by others (Carley, 1997a; Schrodt, 2001). Why would knowledge about this relationship matter? Let's assume somebody provides a new or improved algorithm that leads to a statistically significant increase in REX accuracy. This would be a substantial contribution from an NLP point of view. However, this solution does not tell us anything about what changes we could expect in the shape and properties of the retrieved network. If the changes in network characteristics were also significant, or maybe even larger than the changes in REX accuracy, the need for more accurate REX solutions would be further substantiated, and success in achieving this goal would advance both REX and network analysis. If, however, the impact on the network was insignificant, further investing in improving REX accuracy rates would not be worthwhile from just a network analysis perspective.

In this thesis, I address both of the shortcomings identified and described above in 2.2.1 and 2.2.2, and contribute to a more comprehensive understanding of REX accuracy by raising the following research question:

> *Overall research question:*
> *How much of the variation in a) the structure and properties of network data extracted from texts and b) the results from analyzing these relational data are due to decisions made during the REX process?*

This question is further detailed in the methods section of this chapter. Ultimately, what is needed to answer the raised question in a comprehensive fashion is a knowledge base of method-induced biases and error propagation effects for REX that everybody can draw from when applying or developing such methods and tools. With this thesis, I get work in this direction started by investigating the impact of choices for selected and widely used text coding techniques on network data and analysis results.

Who cares about the outcome of this work? Even though most REX methods have been developed for specific domains and corpora, many of these methods share a large portion of routines for pre-processing and node and edge extraction. I argue that a better understanding of the impact of error propagation and the robustness of REX methods towards these errors contributes to a greater comparability and generalizability of the respective methods. Such knowledge would also provide developers and end-users of REX tools with greater transparency and control over complex, multi-stage analysis processes. Furthermore, a more precise understanding of the relationship between choices made for REX and the robustness of network data and analysis results towards these effects can help end-users to draw valid and reasonable conclusions from their work and work by others. Furthermore, engineers can take this knowledge into account when integrating REX solutions with network analysis technologies. Finally, an answer to the raised research questions is particularly relevant when network data are hard to validate. The knowledge gained with this study can help to weight or factor out effects induced by methodological choices.

2.3 Design of Study and Methodology

How to determine the impact of methodological choices for REX on network data and analysis results? Two strategies are possible: First, one could conduct a series of user studies to observe the coding choices that people make, and then ask human subjects about the conclusions they draw from interpreting the analysis results. The advantage with this approach is that it allows for experimenting with currently relevant domains and genres. However, as outlined in section 2.2, collecting a sufficiently large dataset that allows for drawing generalizable conclusions this way is a costly, long-term process. Alternatively, we could use previously generated and validated datasets. This strategy offers various advantages: it is more cost efficient, does not involve additional reliability tests for human coding, and allows me to focus on the core of the given research question, i.e. the isolation of the impact of user choices on network data. The disadvantage with this strategy is that it limits us to existing datasets, including their constraints. Based on this comparison of strategies, I decided to use the second approach. More specifically, I herein determine the impact of selected methodological choices for REX and the robustness of network data towards these choices by employing the following procedure:

1. Identify a set of relevant methodological choices (sections 2.4 and 2.5).
2. Identify datasets that allow for testing the impact of these choices (section 2.6).
3. Conduct a series of controlled experiments in order to determine the impact of these choices while holding all other choices and factors constant (section 2.7).

2.4 Reference Resolution: Background and Research Questions

Reference Resolution is a widely used pre-processing technique in information extraction and relation extraction. This technique identifies the entity that a referring expression refers to (Hobbs, 1979; Sidner, 1979). For practical applications this means that the various instances and mentions of unique entities, including pronouns, spelling variations, abbreviations, acronyms and repetitions, are identified and consistently associated with or converted into a unique key identifier per entity.

Reference resolution comprises two tasks: anaphora resolution and coreference resolution. The goal with anaphora resolution is to identify the antecedent *A* that an anaphoric expression, also known as anaphor, *B* refers to (Sidner, 1979). Typically, *A* is a noun phrase and precedes *B*, which usually is a pronoun, in the text. *A* is only considered to be an antecedent of *B* if *A* is required for resolving *B*. Thus, the relationship between *A* and *B* is non-symmetric, non-reflexive, and non-transitive (Deemter & Kibble, 2000). The goal with coreference resolution is to identify all of the entities that are mentions of the same referent *C* (Hobbs, 1979). These referring expressions are typically noun phrases. Entity *C* may or may not be explicitly mentioned in the text data. Entities *A* and *B* are only considered to be co-referents if they both unambiguously represent entity *C*, such that A=C and B=C. Therefore, coreferences are symmetric, reflexive, and transitive equivalence relationships (Deemter & Kibble, 2000).

How do anaphora resolution (AR) and coreference resolution (CR) relate to each other? If an anaphor *B* and its antecedent *A* refer to the same entity, *A* and *B* are coreferential. However, there is no deterministic or set-theoretic relationship between AR and CR, i.e. an anaphoric and a coreferential relation may overlap, but not all cases of AR are also cases of CR and vice versa. Another difference between AR and CR is that for resolving a given *B*, in AR, *A* has to be interpreted within the context of the text in which both phrases occur, while in CR, interpreting *A* is not required for testing which entity *C* some entity *B* is identical to. For example, in the phrase "Barack Obama, the President and Nobel Peace Prize winner…", both mentions of a person refer to the real-world entity *C* = "Barack Obama", but an interpretation of entity *A* ("President") is not required for resolving entity *B* ("Nobel Peace Prize winner"). In contrast to that, in the phrase "Obama ran for president in 2008. In 2010, he won the Nobel Peace Prize", resolving the referential expression *B* = "he" with the antecedent *A* = "Obama" requires an interpretation of the text preceding entity *B*.

How is reference resolution (RR) relevant for REX? Both, AR and CR, are normalization and deduplication techniques that are commonly used as pre-processing steps when performing entity extraction and relation extraction. I use the terms *entity* and *node* interchangeably in this chapter since the set of entities contained in a corpus also represents the set of nodes which can be linked into edges. In this context, AR is used to convert pronouns into the respective non-pronominal entities that the pronouns refer to. CR is applied to map multiple instances of an entity to one unique, non-pronominal identifier, and to associate co-referring entities with each other. Taking these effects together, RR can impact the identity, literal mention, i.e. spelling, and weight of nodes and edges. Since there is insufficient knowledge about the impact of RR techniques on network data, I investigate these impacts in this chapter. Furthermore, I argue that the insights gained from this study complement prior knowledge about the effects of deduplication and consolidation of records in relational data, e.g. in relational databases (Bhattacharya & Getoor, 2007; Culotta & McCallum, 2005).

What impact can RR have on network data? Both, AR and CR, can increase the number of mentions per unique entity. This cumulative sum is typically represented as the node weight in network data. While AR does not alter the number of unique named entities, CR potentially reduces this number. Also, while AR is the main strategy for reducing the number of pronouns, CR can also lead to this effect if a set of otherwise unresolved pronouns are identified as being co-referring to each other. Table 3 summarizes these possible effects. Cells labeled as "yes" in Table 3 represent the desired outcome of performing RR.

Table 3: Applicability and impact of reference resolution methods

Case	Type of entity		Applicability of Reference Resolution methods		Potential impact on unique entities (names or nominals, not pronouns)	
	Name or Nominal	Pronoun	Anaphora Resolution	Coreference Resolution	Number	Weight of entities impacted
1	N=1	0	not applicable	not applicable	n.a.	n.a.
2	0	N=1	not applicable	not possible	n.a.	n.a.
3	N>1	0	not applicable	yes	decrease	increase
4	0	N>1	not possible	†	none*	none**
5	N=1	N >= 1	yes	†	none	increase
6	N>1	N >= 1	yes	yes	decrease	increase

† Only among pronouns if number of pronouns > 1

* Decrease of number of distinct pronouns possible

** Increase of weight of unique pronouns

For links, the resolution of anaphoric nodes does not change the link weight. If however two nodes *A* and *B* in a link are coreferences of two nodes *C* and *D* in another link such that *A*=*C* and *B*=*D* or *A*=*D* and *B*=*C*, these two links can be merged into one link while increasing the link weight by one. If further links are merged into this link, the link weight is increased accordingly.

In summary, conducting AR and CR on the entity level is a precondition for impacts of RR on the levels of relational data and network analysis results.

In summary, RR can have the following impact on network data: AR decreases the number of pronominal entities. CR decreases the number of unassociated (not in the sense of unlinked) entities and relations. As a result, both AR and CR can lead to an increase in the number of mentions of unique, non-pronominal entities. If these entities appear as nodes in a network, including isolated nodes, the weight of nodes and links can get increased, and the number of links can get decreased. Combining AR and CR might be more effective in achieving these effects than either technique alone.

Current RR techniques achieve accuracy rates of less than 100%, and no algorithm (or human) might ever return perfectly correct reference resolution results. Actual accuracy rates for RR strongly depend on the applied resolution method, dataset, and evaluation metric. Table 4 gives an overview on performance results for publicly available, state of the art RR technologies; showing that accuracy rates are about 80% and more for AR, and about 70% and higher for CR. The top scoring techniques are based on supervised machine learning methods. In this study, I also simulate the introduction of typical errors into ground truth data in order to understand how much change in RR accuracy rates leads to what changes in network properties.

Table 4: Selection of accuracy rates for Reference Resolution

System	RR	Training data	Evaluation Metric	Recall	Pre-cision	F
Reconcile (Stoyanov et al.)	CR	ACE5	B cubed	55	65	60
Illinois Coreference Package (Bengtson & Roth, 2008), Stanford Deterministic Coreference Resolution System (Raghunathan et al., 2010)	CR, AR and CR	ACE4	B cubed	75	88	81
SemEval2010 (English, information: open, annotation: gold) various participants (Recasens et al.)	CR	SemEval OntoNotes	B cubed	75-85	78-97	82-85
BART (Versley et al., 2008)	AR, CR	ACE2	n.a., B cubed?	55	78	64

This part of the study is driven by the following research question:

> *Overall research question:*
> *What impact does reference resolution have on network data and network analysis results?*

Both of these impacts are referred to as "effects" in the more precise formulation of research questions below. As already explained, both AR and CR can lead to an increase in the number of mentions per unique, non-pronominal entity and the weight of nodes and links, and a decrease in the number of unique links. Since the goal with this project is to understand the impact of RR on network data, I am asking the same research question on the level of entities, links, and network analysis results. Starting from the outlined relationship between RR and network analysis and the logic and functioning of RR techniques, I address the following research questions:

Research question 1:
How large are these effects on the entity level?
Which routine, AR or CR, is more effective in achieving these effects?
Is combining AR and CR more effective than either technique alone?

Answers to the first research question are relevant when conducting NER and content analysis, and for preparing nodes for the construction of network data, for example.

Research question 2:
How large are these effects on the link level?
Which routine, AR or CR, is more effective in achieving these effects?
Is combining AR and CR more effective than either technique alone?

Research question 3:
How large are these effects on the network level?
Which routine, AR or CR, is more effective in achieving these effects?
Is combining AR and CR more effective than either technique alone?

Answering research questions one to three is relevant when performing relation extraction.

Research question 4:
How much change in network properties in due to increases in accuracy rates for AR and CR?

Answers this research question is relevant when selecting a RR technique that is appropriate given the type of network analysis that one plans to conduct.

2.5 Windowing: Background and Research Questions

Once nodes have been identified via entity extraction or some alternative technique, they can be linked into edges in order to construct network data. For this purpose, a variety of approaches have been developed, which exploit lexical (Gerner et al., 1994), semantic (Woods, 1975), syntactic (D. Roth & Yih, 2007), logical (Berners-Lee, Hendler, & Lassila, 2001; Woods, 1975), ontological and taxonomic (Fellbaum, 1998), proximal (Danowski, 1993) and statistical information from text data. A summary of the main families of methods that use these link formation approaches is provided in Table 52. For a more detailed review see also Diesner & Carley (2010c).

Especially in the domain of extracting word networks from texts, which is sometimes also referred to as network text analysis, a commonly used link formation approach is windowing (Carley, 1993; Danowski, 1993). Windowing is a proximity based approach that basically links all entities within a user-defined portion of the text data into edges. Parameters of the window are the chunk of the text input, e.g. sentences or paragraphs, and the number of adjacent words (window size). With some windowing approaches, all identified entities within each window are linked together; forming complete cliques or chains of words (Corman et al., 2002; Gerner et al., 1994). With other approaches, only connections between certain types of nodes (links defined over node types) or nodes that have a specific relationship with each other, such as certain syntactic relationships, are permitted.

The advantages with windowing are that this technique is easy to implement, to adopt for new domains, and to comprehend for end users. These reasons might explain the frequent use of this approach for practical applications. The main critique[2] of windowing is that it is fairly arbitrary and not grounded in theory or any assumption about the production and comprehension of texts (Corman et al., 2002). Moreover, there are hardly any empirical studies of appropriate window sizes given certain domains, datasets, etc., which could provide guidance when selecting a suitable window. I address this gap in research by raising the following research questions:

Research question 5:
What window size do human experts use when identifying relations in text data?
Does this typical window size differ depending on the type of data or relations?

Research question 6:
What window size is needed to capture the vast majority of links in text data?
Does this window size differ depending on the type of data or relations?

Research question 7:

[2] One critique that we have often received on papers that we had submitted and where we used text coding in AutoMap was that the choice of a certain window size was not well justified. One goal with this project is to harness this point of critique.

What error rate, i.e. amount of wrongfully identified links (false positives) and missed links (false negatives), can be expected when applying a specific window size?
Does this error rate differ depending on the type of data or relations?

2.6 Data

For this project, I do not perform reference resolution or windowing manually or algorithmically, but work with sizable datasets that trained human coders have annotated for these tasks. These datasets are assumed to be gold-standard, ground truth data, for which the intercoder-reliability and annotation quality have been previously validated (Jurafsky & Martin, 2009). Using these datasets allows me to make non-probabilistic statements about the impact of the investigated techniques; thus providing an empirically grounded benchmark for the impact of reference resolution techniques and windowing on relational data. Table 5 provides an overview of these datasets and compares them along a few dimensions. These dimensions are relevant for choosing appropriate datasets for the studies presented herein, and show what types of data my findings can reasonably be assumed to generalize to. Table 153 in the Appendix lists the full name and provider ID for each of these datasets.

Table 5: Overview on eligible datasets for information extraction and relation extraction studies in chapters 3 and 4*

Short name	**Full name**	**Enti-ties**	**Relati ons**	**Co-Ref.**	**Ana-phora**	**Genre ****	**Size**	**Year *****	**Used in thesis**
MUC 6	(Chinchor & Sundheim, 2003)	x		x	x (only if coref)	nw (WSJ)	318 articles	1986-1994, 2003	no
MUC 7	(Chinchor, 2001)	x	x	x	x (only if coref)	nw (NYT)	225 articles	1996, 2001	no
ACE 2	(A. Mitchell et al., 2004)	x	x	x	x	news, nw, bcn, ms	518 files	1998, 2003	Ref. Res. (chapter 3)
TIDES 2003	(A. Mitchell et al., 2004)	x	x	x	x	nw, bcn, sp, ms	252 files	2000, 2003	no
ACE 2004	(A. Mitchell, Strassel, Huang, & Zakhary, 2005)	x	x	x	x	nw, bcn, ms	599 files	2000, 2005	no
ACE 2005	(Walker et al., 2006)	x	x	x	x	nw, bcn, bcc, ng, weblogs, ms	599 files	2000-2003, 2006	Ref. Res. and Windowing (chapter 3)

reACE	(Hachey, Grover, & Tobin, 2006)	x	x	x	x	ACE 2004, ACE 2005, BioInfer	900 files (estimate)	2000-2006, 2011	no
BBN	(Weischedel & Brunstein, 2005)	x			x	nw (WSJ)	2454 articles	1989, 2005	Entity Extraction (chapter 4)
Sem Eval 2010-8	(Hendrickx, Kim, Kozareva, & Nakov, 2009)	x (unty-ped)	x			from the web	10718 examples	n.a.	Windowing (chapter 3)
Onto Notes 4	(Weischedel et al., 2011)	x		x		nw, bcn, bcc, ng, web data, ms	353 files (estimate)	2006, 2011	no
Sem Eval 2010-1	(Recasens et al.)	x		x		see OntoNotes 4	353 files	2006, 2010	no
NYT AC	(Sandhaus, 2008)	x		x		nw (NYT)	1.5 Mio. Articles	1987-2007, 2008	no
CoNLL 2003	(CoNLL-2003, 2003)	x				nw, Reuters corpus	1393 files	1996-1997, 2000	no

* only English text data considered for this thesis

** nw = newswire, bcc = broadcast conversations, bcn = broadcast news, sp = speech, ng = newgroups, ms = from multiple sources (not genres, but different news paper for example)

***first number: source (English), second number: data source provider

For the reference resolution project, data are needed in which sufficiently large amounts of anaphoric relations, coreferential relations, and other types of relations between entities are annotated. Eligible datasets are MUC and ACE (incl. TIDES and reACE) (Table 5). In MUC, however, relations are restricted to specific types of links between entities and organizations only, and the total number of marked up relations (N = 800) is lower by factor of ten than in ACE (Table 6). For these reasons, MUC was not selected for this project. Given that all ACE datasets would be appropriate for this project based on their size and breadth of types of relations considered, I choose to use the oldest (ACE2) and newest (ACE5) one listed in Table 5. The reason for this decision is that it allows for testing whether findings are robust over time (the difference in publishing date of the articles in these corpora is five years). Furthermore, ACE 2 and ACE 5 are similar in the amount and type of annotated relations, thus enabling reasonable comparisons (Table 6). They also overlap in genre - both cover printed and spoken news data – which again facilitates comparisons across time. In addition to that, ACE covers three additional genres, namely blogs, online discussion groups, and telephone conversations, which allows for testing differences between genres.

For the windowing project, data are needed in which large numbers of instances of different types of relationships are marked up so that the robustness of findings across differences types of relations can be assessed. Table 6 provides a comparison of the types of relations per corpus. In order to provide consistency in this chapter, I choose to use ACE5 for this project again. From all of the various ACE datasets, ACE5 offers the greatest variety of genres and types of relations to analyze (syntactic relations, semantic relations, relations defined over node types). Since I am aiming for generalizability of the findings from this study, it seemed important to consider different points of comparison, which rules out ACE2 because the annotation guideless for establishing relations are very similar for ACE2 and ACE5 (in fact, they were developed over time from the same baseline). The only dataset that fulfills the outlined criteria for a suitable dataset and provides some different types of data and relations as ACE5 does is SemEval, which was therefore was chosen as the second dataset for the windowing project.

Table 6: Comparison of relations in datasets

Size of dataset and comments	Types of relations considered
MUC 7 N = 800 relations between entities and organizations only	1. Employee of 2. Product of 3. Location of
ACE 2, TIDES N = 8,127 all defined over entity types further classifications: class: explicit, implicit	1. Role: employment (management, general staff), other (member, owner, founder, client, affiliate-partner, citizen-of, other) 2. Part: subsidiary, part-of, other 3. At: located, based in, residence 4. Near: relative location 5. Social: personal (parent, sibling, spouse, grandparent, other relative, other personal), professional (associate, other profess.)
ACE 2004 some defined over entity types	1. Physical: located, near, part whole 2. Personal/Social: business, family, other 3. Employment/Membership/Subsidiary: employ-exec(s), employ-staff, employ-undetermined, member of group, subsidiary, partner, other 4. Agent-Artifact: user/owner, inventor/ manufacturer, other 5. Person-Organization: ethnic, ideology, other 6. GPE Affiliation: citizen/resident, based in, other 7. Discourse
ACE 2005 N = 8,738 all defined over entity types further classifications: syntactic relation, modality, tense	1. Physical: located, near 2. Part whole: geographical, subsidiary, artifact 3. Personal/ social: business, family, lasting-personal 4. ORG Affiliations: employment, ownership, founder, student-alum, sports-affiliation, investor-shareholder, membership 5. Agent-Artifact: user-owner- inventor-manufacturer

	6. Gen-Affiliation: citizen-resident-religion-ethnicity, org-location-
SemEval 2010-8 N = 10,717 not defined over entity types, entity types not labeled	1. Cause-Effect 2. Component-Whole 3. Content-Container 4. Entity-Destination 5. Entity-Origin 6. Instrument-Agency 7. Member-Collection 8. Message-Topic 9. Other 10. Product-Producer

2.6.1 Preparation Datasets for Experiments

The datasets selected for this project use different ways of marking up entities, relations, and other text properties that are needed herein. Therefore, I built a parser for each datasets in order to extract the required information. I briefly describe the details on this process to the minimum extent needed for ensuring the reproducibility of my results.

In ACE, the text files are marked up in SGML format. These SGML files contain only the raw texts and meta-data, such as the source and release date of an article. The information on entities and relations is specified in XML files. In these XML files, entities and relations have a head (key word or key phrase) and an extent (typically a nominal phrase). The mapping from the XML files to the text files is realized through position numbers. This numbering pauses at SGML tags within the file body. I herein consider elements of the types "entity" and "timex" as entities. Entities of the type "timex" are included because they represent instances of the "time" class in the meta-network model. The meta-network model is a theoretically grounded model of relevant classes of entities and links in socio-technical networks (for a more detailed description of this model see section 3.2.4). The mentions of entities in the data are categorized as names, nominals or pronouns. Pronouns include terms like "one", "some" and "there".

In ACE, the "smallest or closest possible relation" is tagged, typically on the sentence level (Linguistic_Data_Consortium, 2005). A few relations span across sentences. In general, analyzing gold standard information about window sizes across sentences would contribute new knowledge, but since this option violates the preferred norms in ACE, I did not further explore this path.

Relations are coded as follows in ACE: if two entity mentions *C* and *D*, which are instances of a pair of nodes that involves entity mentions *A* and *B* such that *A*=*C* and *B*=*D* or *A*=*D* and *B*=*C* are identified to form the same type of relationship, the respective relationship is annotated to have multiple mentions (in this case two). If the type of relationships is different, the relations are marked up as different relations. In order to identify the impact of CR on relational data, I deviate from this notion of link identity by using the following operationalization: any two links that are marked up in a given text are identical if both entity mentions contained in one link map to the same entities as the entity mentions in another link, regardless of directionality.

Finally, ACE2 contains 20 redundant relations (same type of relationship between identical nodes at same text position), which I removed in order to deduplicated these links. ACE 2005 contains four relations where the head of both nodes were identical (same token at same position in same file). I disregarded these four relations for the entity level analysis since they would dilute the coreference resolution results (even though the impact is minimal), but kept them for analyses on the relation and network level.

2.6.2 Selection of Relevant Aspects of Relational Data for Analysis

The ACE data have been previously used by others to develop and validate cutting-edge reference resolution techniques (Doddington et al., 2004). Both selected ACE datasets allow for studying the impact of reference resolution and windowing on multiple aspects of relational data. These aspects include the type or genre of the data, different node classes, and the type of relations, such as different semantic relations. Therefore, a selection of aspects that are relevant for the context of this thesis is necessary. For the RR project, I have already explained why analyses will be conducted on the level of nodes, links, and network data. For windowing, this choice is inapplicable as windowing only impacts the network data level; thus analysis will be conducted on that level only. Moreover, for the windowing study, multiple aspects of relations that are relevant for network analysis are being considered, namely the genre of the data and the type of nodes and links. Given that for the RR project, I decided to conduct analysis on the entity, link and network data level, this comprehensive scope needed to be limited. For practical text analysis projects, one unanswered question that we often face is the following (Carley et al., 2007; Dabbish et al., 2011): What coding choices would be appropriate for some specific type of data? For example, when analyzing well-formed news data, different choices and techniques might be appropriate than when analyzing data from social networking platforms, which often feature a more informal orthography and grammar. Therefore, I decided to test the impact of RR techniques on different genres. Table 7 compares the genres available in ACE with respect to the number of agents involved in producing a piece of text data, whether the text comes from written or spoken language, and the level of formality. ACE2 covers the first two genres presented in Table 7, and ACE5 covers all three of them.

Table 7: Characteristics of data per genre (ACE)

Levels of compare-son between genres		Newswire	Broadcast news	Broadcast conversat.	Telephone	Usenet	Weblogs
Number of agents	Conversation			x		x	x
	Dialogue				x		x
	Monologue	x	x				x
Mode	Written	x				x	x
	Spoken		x	x	x		
Style	Formal	x	x	x			
	Informal				x	x	x

2.7 Results

The presented results are based on the judgment of trained people who aimed to deliver the best reference resolution and windowing results that humans can possibly provide. Therefore, my findings report on the upper bound of the impact of highly accurate reference resolution on entity extraction, relation extraction, and network analysis.

2.7.1 Reference Resolution

In general, two strategies are available for analyzing the impact of reference resolution on nodes, edges and network data: first, one could use only the entities that are involved in relations. Second, the full set of entities marked up in the corpus could be used. I chose the second strategy for the following reasons: first, even if an entity is not involved in a link, it might still show up as an isolated node in a graph. In fact, in network analysis, people consider isolates for certain analyses, e.g. in the context of organizational networks and covert networks (Klerks, 2001). The metric of "connectedness" was developed to measure the ratio of isolates to connected nodes in a network (Wasserman & Faust, 1994). Second, whether a node is connected into a link or not strongly depends on the mechanism for link creation; with some techniques being more inclusive than others (see sections 2.5 and 3.2.3 for details on methods for link creation). Third, it is possible that an isolated node gets mapped onto another, already connected node via reference resolution techniques such that the weight of the linked node is increased. In order to provide a comprehensive understanding of the upper bound of the impact of reference resolution on relational data, I decided to analyze the entire set if entities.

The distribution of names, nominals and pronouns per genre (Figure 3, Figure 4[3]) shows that written news data data are atypical in their frequent use of names and less frequent use of

[3] Note that Figure 3 represents the same information as Figure 4 and Figure 5 together, but since there are more genres in ACE5 than in ACE2 (Figure 4, Figure 5), I had to split up the information into two graphics to avoid overcrowding.

pronouns. Therefore, in comparison across genres, AR seems potentially least effective for news data, and can have a higher impact on all accounts of informal writing and spoken language, especially telephone conversations. The information presented in Figure 3 and Figure 4 also shows that when working with news data only (ACE2), a biased perception of the distribution of entity types emerges, which could underestimate the role of pronouns and thus AR, and overestimate the weight of names and nominals and thus the impact of CR.

The ratio of first mentions of unique entities to additional entity mentions is fairly similar across genres (Figure 3, Figure 5). Repeated references to previously introduced entities are most prevalent among pronouns: on average, about 2/3 of pronoun mentions are back-references. This further stresses the importance of AR. Also, this finding suggest that while pronouns are typically thought of as candidates for AR, it could be worthwhile to also apply CR to them, especially if no name or nominal is available that could serve as an antecedent. The ratio of first mentions to repetitions is inverse for nominals (over 2/3 are unique, first time mentions). For names, well over half of all mentions are references to previously introduced entities.

Figure 3: Distribution of entity types (mentions) per genre (ACE2)

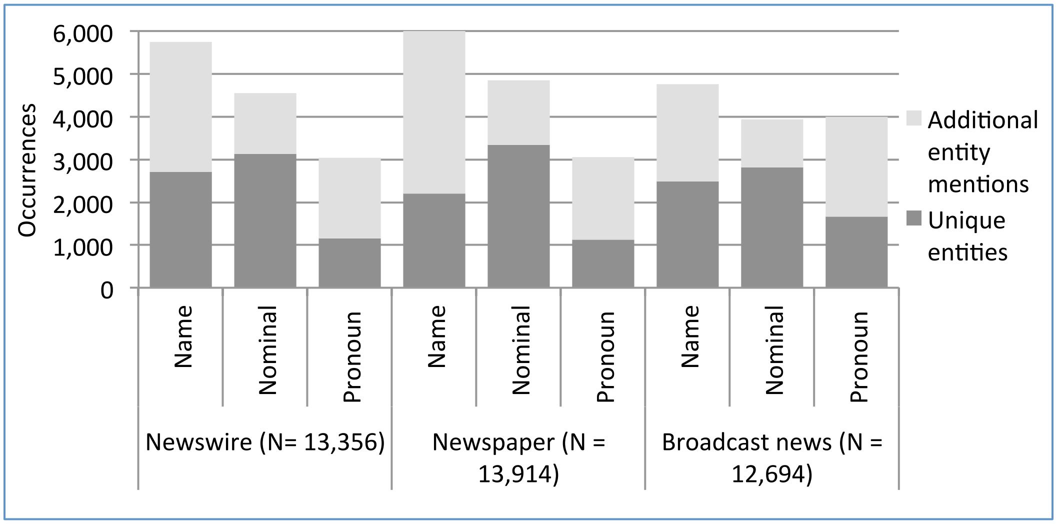


Figure 4: Distribution of entity types (mentions) per genre (ACE5)

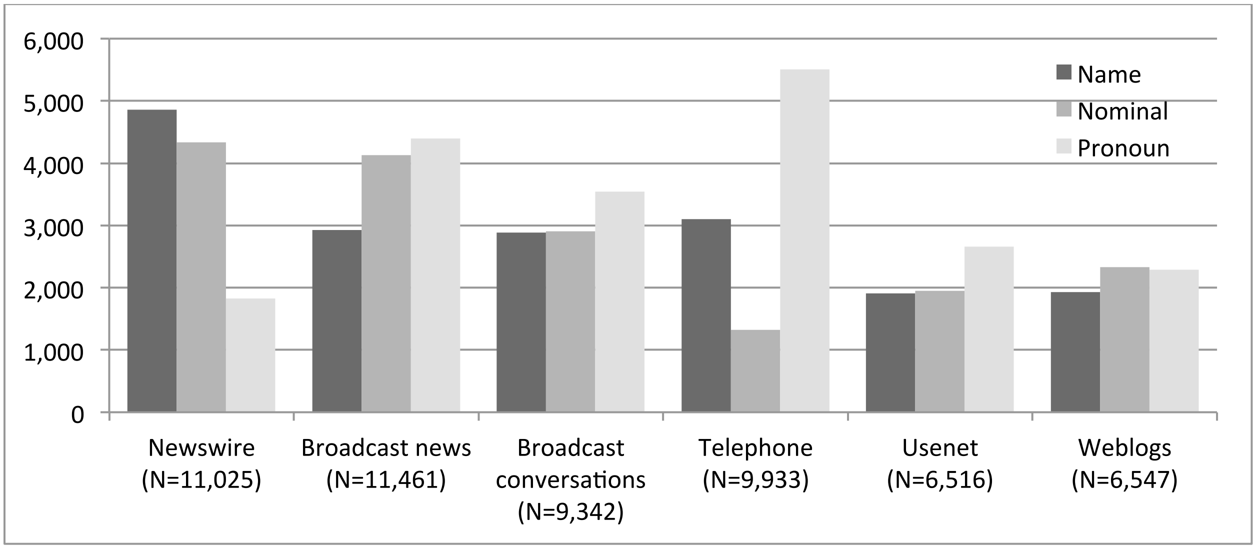
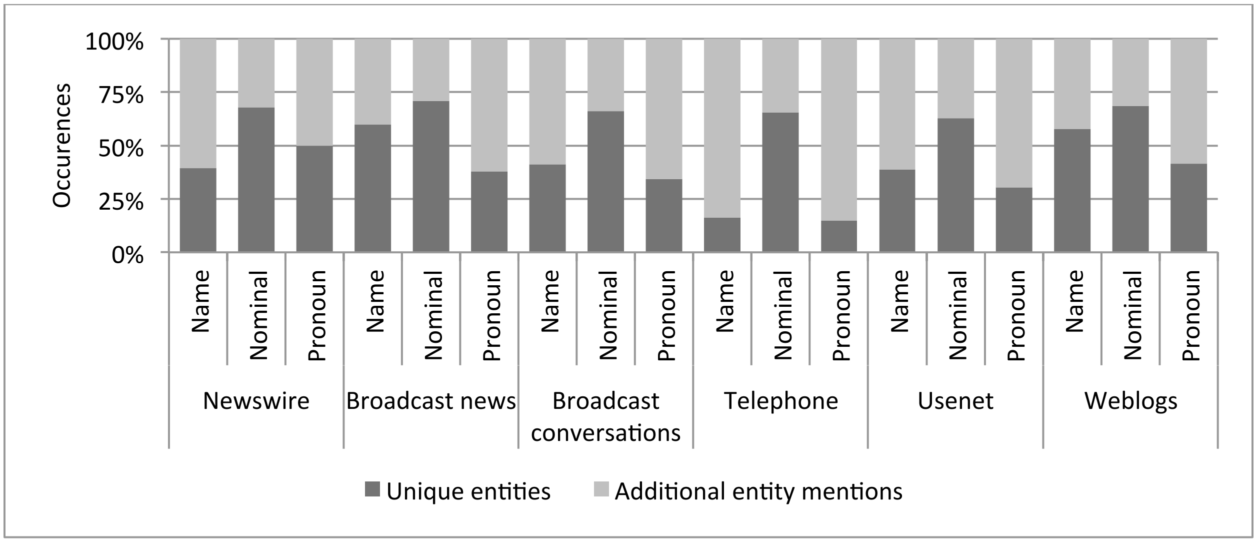


Figure 5: Ratio of unique entities and their additional mentions by entity type and genre (ACE5)



2.7.1.1 Impact of Reference Resolution on Entities

Depending on the genre, about 60% and more of all entity mentions are subject to reference resolution (Figure 6, Figure 7). More specifically, pronouns account for roughly 40% of all entities mentions (about 20% for newswire and newspaper data, more than 50% for telephone data). These entities are subject to AR. Depending on the genre, additional mentions of unique names and nominals constitute another 20% to 30% of the data (40% to 50% for newswire and newspaper data). These entities are subject to CR. Given the distributions of entity types, theoretically, AR can have a bigger impact than CR on altering the identity and weight of nodes for six of the nine genres considered.

Figure 6: Entity mentions that are subject to change or not (ACE2)

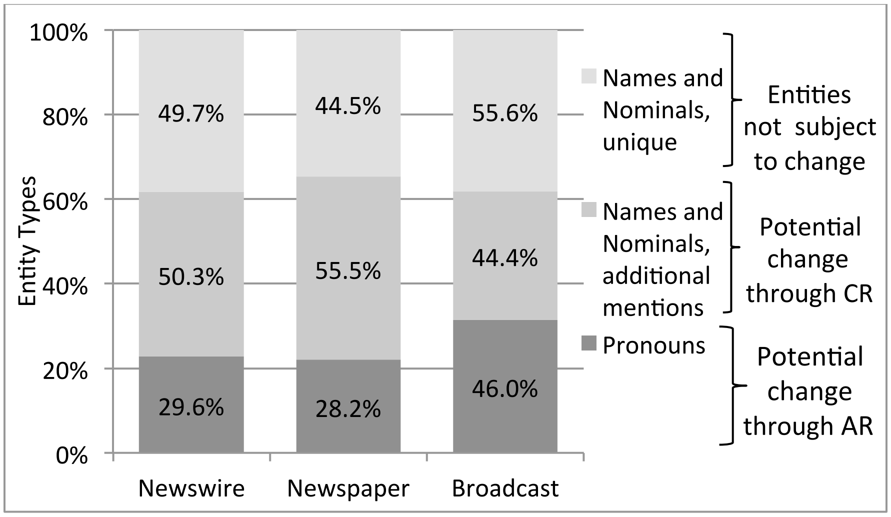
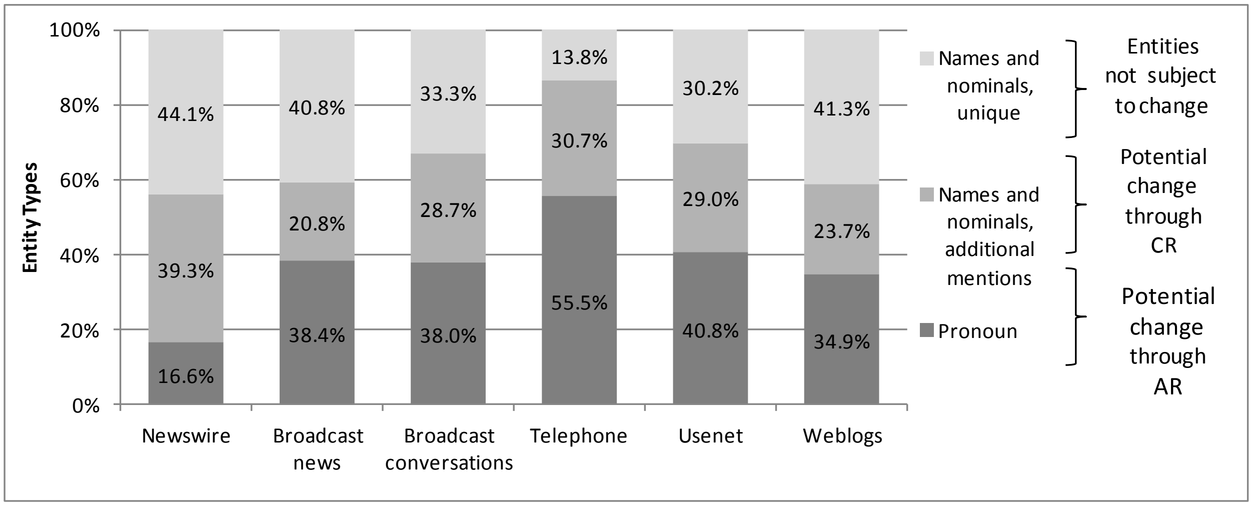


Figure 7: Entity mentions that are subject to change or not (ACE5)



In this project, anaphora are considered as irresolvable via AR if all mentions of a pronoun are also pronouns. The results for AR show that for all genres, the majority of pronouns can be resolved (between 67% and 86% of pronoun mentions), resolution rates are higher for written texts than for spoken language, and the highest resolution rates are achieved where the ratio of pronouns is lowest (newswire and newspaper data, 83% to 85%) (Table 8, Table 9). I speculate that for transcripts of spoken language, AR is complicated by the fact that these data have proportionally more pronouns to begin with, and therefore a smaller pool of names and nominals is available to associate the pronouns with. Most anaphora are resolved by both names and nominals. This indicates that conducting CR after AR is another crucial step. Nominals are slightly more effective in leading to this effect than names. This suggests that the availability of entities that are not referred to by a name, such as role descriptors, facilitates the RR process, which is important with respect to the selection of nodes classes for entity extraction in section 3.2.5. More than 65% of all irresolvable pronouns (46% for telephone data) are pronouns that have only one mention. They will remain in the data the way they are; accounting for 2% to 14% of all entities per genre. Unresolved pronouns that have multiple mentions can be grouped into clusters per unique entity. Single mentions of names and nominals can serve as antecedents for AR. On average, applying CR to unresolved anaphora helps to group more than 2/3 of all pronouns that cannot be resolved via AR into clusters that refer to the same entity (Table 12, Table 13).

Table 8: Results for anaphora resolution per genre (ACE2)

	Newswire	Newspaper	Broadcast news
	Unique entities		
Resolved by name(s) only	15.2%	13.6%	17.9%
Resolved by nominal(s) only	28.5%	30.2%	23.9%
Res. by both only	26.6%	29.1%	15.9%
Sum resolved	70.3%	72.9%	57.7%
Unresolved	29.7%	27.1%	42.3%
Single mentions in unres.	78.3%	76.9%	65.6%
	Entity mentions (including first mention)		
Resolved by name(s) only	12.1%	10.8%	15.7%
Resolved by nominal(s) only	19.1%	17.5%	18.5%
Resolved by both only nominal(s)	51.8%	57.0%	32.7%
Sum resolved	82.9%	85.4%	66.9%
Unresolved	17.1%	14.6%	33.1%
Resolved anaphora in corpus	18.9%	18.8%	21.3%
Irresolvable anaphora in corpus	3.9%	3.2%	10.4%

Table 9: Results for anaphora resolution per genre (ACE5)

	Newswire	Broadcast news	Broadcast conversat.	Telephone	Usenet	Weblogs
	Unique entities					
Resolved by name(s) only	9.3%	13.3%	14.2%	16.5%	23.0%	18.1%
Resolved by nominal(s) only	32.5%	28.5%	31.0%	26.4%	27.7%	34.5%
Res. by both only	34.8%	17.2%	17.7%	13.4%	10.3%	21.5%
Sum resolved	76.5%	59.1%	62.9%	56.3%	61.0%	74.1%
Unresolved	23.5%	40.9%	37.1%	43.7%	39.0%	25.9%
Single mentions in unres.	84.7%	62.7%	65.3%	46.3%	65.4%	70.3%
	Entity mentions (including first mention)					
Resolved by name(s) only	11.1%	12.1%	14.2%	34.8%	28.0%	25.4%
Resolved by nominal(s) only	23.9%	23.6%	25.1%	13.1%	25.7%	21.6%
Resolved by both only	50.7%	33.1%	34.0%	26.1%	22.6%	33.1%
Sum resolved	85.8%	68.8%	73.3%	74.0%	76.4%	80.1%
Unresolved	14.2%	31.2%	26.7%	26.0%	23.6%	19.9%
Resolved anaph. in corpus	14.2%	26.4%	27.9%	41.1%	31.1%	28.0%
Irres. anaphora in corpus	2.4%	12.0%	10.1%	14.4%	9.6%	6.9%

The results for CR show that about 30% to 40% (17% for telephone) of all names and nominals together are single mentions such that CR does not apply (Table 10, Table 11). Overall, most co-referencing happens via a mixture of names and nominals. This ratio of single mentions is about twice as high for nominals than for names, which does not reflect the distribution of entities in the data (there are typically more or as many names than nominals.

Table 10: Results for co-reference resolution by genre (ACE2)

	Newswire	Newspaper	Broadcast
	Unique entities		
Single Names	27.4%	21.5%	27.5%
Single Nominals	38.6%	46.5%	41.2%
Name co-ref. by Name	11.5%	9.4%	14.1%
Nominal co-ref. by Nom.	8.3%	8.0%	7.4%
Mixed co-referencing	14.2%	14.6%	9.8%
Sum singles	66.0%	68.0%	68.6%
Sum co-referenced	34.0%	32.0%	31.4%
	Entity mentions (including first mention)		
Single Name	13.6%	9.6%	15.2%
Single Nominal	19.2%	20.7%	22.8%
Name co-ref. by Name	19.1%	15.8%	21.9%
Nominal co-ref. by Nom.	12.1%	10.6%	11.0%
Mixed co-referencing	36.0%	43.4%	29.0%
Sum singles	32.8%	30.2%	38.0%
Sum co-referenced	67.2%	69.8%	62.0%
Sum co-ref. in corpus	51.9%	54.4%	42.4%

Table 11: Results for co-reference resolution by genre (ACE5)

	Newswire	Broadcast news	Broadcast conversat.	Telephone	Usenet	Weblogs
	Unique entities					
Single Names	18.9%	22.2%	18.8%	16.3%	21.3%	26.9%
Single Nominals	43.0%	47.4%	45.9%	44.1%	45.9%	43.5%
Name co-ref. by Name	8.4%	7.4%	11.9%	13.5%	12.8%	6.9%
Nominal co-ref. by Nom.	9.9%	10.3%	11.3%	14.8%	12.8%	10.1%
Mixed co-referencing	19.8%	12.6%	12.0%	11.3%	7.3%	12.5%
Sum singles	61.9%	69.6%	64.8%	60.4%	67.1%	70.4%
Sum co-referenced	38.1%	30.4%	35.2%	39.6%	32.9%	29.6%
	Entity mentions (including first mention)					
Single Name	8.4%	13.1%	9.0%	4.5%	9.9%	15.2%
Single Nominal	19.0%	27.9%	22.0%	12.3%	22.3%	24.6%
Name co-ref. by Name	16.9%	11.6%	20.6%	45.5%	22.3%	11.5%
Nominal co-ref. by Nom.	13.5%	16.4%	14.8%	11.5%	19.9%	15.8%
Mixed co-referencing	42.2%	31.0%	33.5%	26.1%	25.5%	32.9%
Sum singles	27.3%	41.0%	31.0%	16.9%	32.3%	39.8%
Sum co-referenced	72.7%	59.0%	69.0%	83.1%	67.7%	60.2%
Sum co-ref. in corpus	36.2%	36.4%	71.5%	37.0%	40.1%	39.2%

Putting the results for AR and CR on the entity level together shows that across genres, the considered reference resolution techniques can alter the identity and weight of at least 70% of all entity mentions (Table 12, Table 13). Entities that are not changed by reference resolution techniques are either irresolvable pronouns (less than 4% of all remaining entities), or names and nominals that are mentioned only once, which might still have been essential for AR (about 15% to 26% of all entities). I had shown that given the raw frequencies of pronouns, names and nominal, AR could have a stronger impact on entities than CR. However, the results indicate that CR contributes more strongly to the desired entity normalization and consolidation effects for all but the telephone data. One explanation for this counterintuitive finding might be the fact that AR increases the set of entities applicable to CR in the first place. Another interesting finding is that CR on pronouns that could not be resolved via AR has a minor yet meaningful impact on the data (less than 1% up to 13% of all entities in the resulting data). Finally, the results show that combining AR and CR is more effective than using either technique alone.

Table 12: Summary of effectiveness of reference resolution techniques by genre (entity mentions, ACE2)

	Reference Resolution technique	Newswire	News-paper	Broadcast news
Anaphora	Resolved with AR	18.9%	18.8%	21.1%
	Resolved with CR	1.9%	1.5%	6.8%
	Unresolved	2.0%	1.7%	3.6%
Names & Nominals	CR	51.9%	54.4%	42.4%
	No CR	25.3%	23.6%	26.1%
Summary	Change through AR	20.8%	20.3%	27.9%
	Change through CR	51.9%	54.4%	42.4%
	Change through RR	72.7%	74.7%	70.3%

Table 13: Summary of effectiveness of reference resolution techniques by genre (entity mentions, ACE5)

Impact on	Reference Resolution technique	News-wire	Broadc. news	Broadc. convers.	Tele-phone	Usenet	Weblogs
Anaphora	Resolved with AR	14.2%	26.4%	27.9%	41.1%	31.1%	28.0%
	Resolved with CR	0.7%	8.2%	7.0%	12.8%	6.5%	4.0%
	Unresolved	1.7%	3.7%	3.2%	1.7%	3.2%	2.9%
Names & Nominals	CR	60.6%	36.4%	42.8%	37.0%	40.1%	39.2%
	No CR	22.8%	25.2%	19.2%	7.5%	19.1%	25.9%
Summary	Change through AR	14.9%	34.7%	34.8%	53.8%	37.6%	32.0%
	Change through CR	60.6%	36.4%	42.8%	37.0%	40.1%	39.2%
	Change through RR	75.5%	71.0%	77.6%	90.8%	77.7%	71.2%

In the original set of all entities, the weight of each distinct entity mention equals one. This deviates a bit from common procedure in practical entity extraction and REX applications, where orthographically identical entities are sometimes considered as the same entity. When applying thesauri in AutoMap, for example, all identically spelled concept – regardless of capitalization – are translated into the same entity with no further word sense disambiguation routines applied. This procedure greatly eases the efforts required for building thesauri, but implies the danger of false positives, e.g. in the case of homographs and heteronyms, and of false negatives, e.g. in the case of synonyms. Does the separation of identical terms from heteronyms matter with respect to entity weights? Mapping entities onto each other not based on spelling, but proper word sense disambiguation as approximated via reference resolution techniques shows that for the unique entities affected by this procedure, the average node weight is increased from 1.0 to 5.1 with AR, to 4.6 with CR, and to 6.0 when using both techniques (Table 14, Table 15). Consequently, a significant portion of the total node weight in the dataset shifts to these entities: using both, AR and CR, causes less than 20% of the unique entities carry more than 75% of the total node weight, while the remaining more than 80% of unique entities carry less than 25% of the total weight. This means that reliable reference resolution helps not only to disambiguate entities, but also to increase and enrich the amount of information available on truly distinct entities. This is particularly valuable when working with sparse networks, and sparseness is common feature of large-scale, real-world networks (Barabási & Albert, 1999).

Table 14: Comparison of impact of reference resolution techniques on entity reduction and node weights (ACE2, averaged across genres)

	Decrease in no. of unique entities (corpus)	Entities impacted by routine			Entities not impacted by routine (node weight = 1)	
		Amount	Total node weight carried	Average node weight	Amount	Total node weight carried
AR	19.56%	8.1%	26.0%	4.01	91.9%	74.0%
CR on pronouns	2.35%	1.0%	3.3%	3.42	99.0%	96.7%
CR	37.72%	19.3%	49.8%	4.13	80.7%	50.2%
AR and CR	59.63%	38.0%	74.9%	4.89	62.0%	25.1%

Table 15: Comparison of impact of reference resolution techniques on entity reduction and node weights (ACE5)

Genre	Decrease in no. of unique entities (corpus)	Entities impacted by routine			Entities not impacted by routine (node weight = 1)	
		Amount	Total node weight carried	Average node weight	Amount	Total node weight carried
	AR					
Newswire	14.2%	6.3%	20.6%	3.2	93.7%	79.4%
Broadcast news	26.4%	8.6%	35.0%	4.1	91.4%	65.0%
Broadcast con.	27.9%	8.2%	36.1%	4.4	91.8%	63.9%
Telephone	41.1%	4.6%	45.7%	9.9	95.4%	54.3%
Usenet	31.1%	7.6%	38.7%	5.1	92.4%	61.3%
Weblogs	27.4%	9.5%	36.9%	3.9	90.5%	63.1%
Average	28.0%	7.5%	35.5%	5.1	92.5%	64.5%
	CR on pronouns					
Newswire	0.4%	0.3%	0.7%	2.4	99.7%	99.3%
Broadcast news	6.0%	2.2%	8.2%	3.7	97.8%	91.8%
Broadcast con.	5.3%	1.7%	7.0%	4.1	98.3%	93.0%
Telephone	10.9%	1.9%	12.8%	6.6	98.1%	87.2%
Usenet	4.8%	1.7%	6.5%	3.9	98.3%	93.5%
Weblogs	1.0%	1.5%	2.5%	1.7	98.5%	97.5%
Average	4.7%	1.6%	6.3%	3.7	98.5%	93.7%
	CR (Names and Nominals)					
Newswire	46.6%	14.0%	60.6%	4.3	86.0%	39.4%
Broadcast news	25.4%	11.0%	36.4%	3.3	89.0%	63.6%
Broadcast con.	32.3%	10.5%	42.8%	4.1	89.5%	57.2%
Telephone	32.1%	4.9%	37.0%	7.5	95.1%	63.0%
Usenet	31.1%	9.1%	40.1%	4.4	90.9%	59.9%
Weblogs	28.3%	10.9%	39.2%	3.6	89.1%	60.8%
Average	32.6%	10.1%	42.7%	4.5	89.9%	57.3%
	AR & CR					

Newswire	61.2%	16.1%	77.4%	4.8	83.9%	22.6%
Broadcast news	57.8%	17.2%	75.0%	4.4	82.8%	25.0%
Broadcast con.	65.4%	15.2%	80.6%	5.3	84.8%	19.4%
Telephone	84.0%	8.3%	92.3%	11.1	91.7%	7.7%
Usenet	67.0%	13.5%	80.5%	5.9	86.5%	19.5%
Weblogs	58.3%	16.9%	75.1%	4.5	83.1%	24.9%
Average	65.6%	14.5%	80.2%	6.0	85.5%	19.9%

2.7.1.2 Impact of Reference Resolution on Links

Not all entities that are retrieved from some text data as potential nodes for networks will be linked into edges. This can be for two reasons: first, some entities are truly not related to any other entities (isolates), but can still be meaningful for network analysis. In the considered data, about 28% (ACE5) to a third (ACE2) of all entity mentions and a little over half of all unique entities (ACE2 and ACE5) do occur in relations. Since over 70% of all entities mentions are impacted by RR, it is seems highly likely that some of the entities occurring in edges are affected by RR. Second, in most ground truth data for REX, relations are mainly annotated within sentences, but not across sentences, paragraphs or documents. Besides the previously mentioned sparseness that has been observed for many real-world networks, these two reasons also contribute to the sparseness of relational data available for studying REX. Consequently, the density of the relational data used herein, which is computed as the number of actual relations over the number of possible relations, is very low across genres (Table 16, Table 17) (Wasserman & Faust, 1994).

The ratio of relations that contain at least one node that is a pronoun is very similar across genres in ACE 2, which are all news coverage (average 16%, Table 16), and varies widely in ACE5 (12% to 70%, Table 17). The following analyses require some definitions: Let's assume that AR on the link level is only successful if all pronominal nodes in a link can be resolved by a name or nominal. This conservative operationalization is referred to as "AR strict" in the following tables, and allows for determining the minimum amount of change that AR can cause on the link level. Using this approach, the AR rate is high and highly similar across genres; about 75%-78% for spoken data and 79% to 85% for written data. Since the rate of links involving pronouns varies per genre, the ratio of links that are altered due to AR ranges from 9% to 52% (Table 16, Table 17). Relaxing the strict operationalization of successful AR on the link level to assuming that AR is successful if at least one pronoun in a link is resolvable marginally increases the AR rate by an average of 0.6% (Table 17: AR relaxed, analysis with this operationalization conducted for ACE5 only). This additional gain is small for the following reason: in addition to the links impacted by the strict operationalization, the relaxed version also affects links in which both nodes are a pronoun. This applies to 6.3% of all links that entail a pronoun, and more than half of them were already completely resolved under the strict AR condition. In the next step, all nodes on which AR was successful become additional candidates for CR.

Per genre, the number of links between only names and nominals (candidates for CR) is very similar within ACE 2 (83% to 85%, Table 16), and again varies strongly in ACE5 (29% to 82%, Table 17) [4]. The ratio of links that gets reduced via CR when multiple links are mapped onto one link ranges from 4% to 12%.

As previously explained, CR can also be applied to anaphora[5]. I have operationalized CR on anaphora for the link level as follows: CR on anaphora is successful if both entity mentions in a link are pronouns, and both pronouns map to the same entities as the entity mentions in another link, which are also anaphora. This effect is much smaller than regular CR on the link level (on average 0.3% in ACE5, Table 17), and smaller than CR on pronouns on the entity level.

On the relational data level, the interaction between AR and CR is as follows: while the relation reduction is entirely due to CR, AR provides a large amount of names and nominals available to CR. Combining AR and CR has a stronger impact on consolidation edges than using either technique alone (last row in Table 16, Table 17): on average, an additional 3% to 4% of all link mentions are reduced. This rate is even higher for telephone and usenet data (not included in average reported in previous sentence), where link reduction rates of 18% to 19% were observed.

Table 16: Results for impact of AR and CR on relational data (ACE2)

RR technique applied	Measure of impact of RR on data	Newswire	Newspaper	Broadcast
none	Number of links	2,884	2,956	2,267
	Number of entity mentions	13,356	13,914	12,694
	Density	0.0032	0.0031	0.0028
AR strict	Links with pronoun	14.8%	16.7%	16.5%
	…, pronoun resolved	76.6%	87.0%	76.1%
	…, resolved in corpus	11.3%	14.5%	12.5%
CR	Links with names and nominals	85.2%	83.3%	83.5%
	…, reduced via CR	4.2%	4.7%	7.5%
AR + CR	Links reduced in corpus	6.5%	7.9%	10.6%

[4] For ACE5, the ratio of links with pronouns and links with names and nominals does not add up to 100% due to the inclusion of entities of type timex in links. These entities are not names, nominals or pronouns.

[5] In ACE2, there were only three links for which CR was possible on pronouns. Since these effects are marginal I disregard them from analysis on the relation data level.

Table 17: Results for impact of AR and CR on relational data (ACE5)

RR technique applied	Measure of impact of RR on data	News-wire	Broadc. news	Broadc. conv.	Tele-phone	Usenet	Web-logs
none	Number of links	2,683	2,016	1,660	746	864	769
	Number of entity mentions	11,025	11,461	9,342	9,933	6,516	6,547
	Density	0.0044	0.0031	0.0038	0.0015	0.0041	0.0036
AR strict	Links with pronoun corpus	11.9%	29.6%	25.7%	69.6%	49.4%	26.9%
	…, pronoun resolved	79.6%	78.4%	76.6%	75.0%	78.9%	84.1%
	…, resolved in corpus	9.4%	23.2%	19.7%	52.1%	39.0%	22.6%
	…, unresolved in corpus	2.4%	6.4%	6.0%	17.4%	10.4%	4.3%
relaxed	…, pronoun resolved	80.8%	80.2%	79.4%	76.9%	80.3%	85.0%
	…, resolved in corpus	9.6%	23.7%	20.4%	53.5%	39.7%	22.9%
	…, unresolved in corpus	2.3%	5.9%	5.3%	16.1%	9.7%	4.0%
CR	Links /w name & nomin.	82.0%	65.2%	71.6%	29.1%	49.0%	70.1%
	…, no CR possible	90.0%	93.6%	88.5%	92.6%	88.7%	93.5%
	…, no CR possible in corpus	73.9%	61.0%	63.3%	26.9%	43.4%	65.5%
	…, reduced via CR	10.0%	6.4%	11.5%	7.4%	11.3%	6.5%
	…, reduced via CR in corpus	8.2%	4.2%	8.3%	2.1%	5.6%	4.6%
	…, reduced via CR on anaphora in corpus	0.0%	0.4%	0.4%	0.5%	0.3%	0.0%
	…, sum reduced in corpus	8.2%	4.6%	8.6%	2.7%	5.9%	4.6%
AR + CR	Links reduced in corpus	10.9%	9.7%	13.4%	19.0%	18.4%	8.6%

Overall, the link normalization and deduplication effects due to RR are less strong on the link level than on the entity level (Table 18: values averaged over genres, Table 19). For example, on the entity level, the average weight of unique entities impacted by both AR and CR increases from 1.0 to 5.5, while on the link level, the average weight of impacted unique relations increases to less than 2.3. Moreover, the results indicate that on the entity level, CR has a stronger impact (average entity reduction rate = 45.0%) than AR (average entity change rate = 30.8%) does, while on the link level, AR (average link change rate = 22.7) is more effective than CR (average link reduction rate = 5.7%).

Table 18: Comparison of impact of reference resolution techniques on link level, averaged over genres (ACE2)

Case	Impact on data					
	Link change rate (AR), link reduction rate (CR, AR & CR)	Entities impacted by routine			Entities not impacted by routine (node weight = 1)	
		Amount	Total node weight carried	Average node weight	Amount	Total node weight carried
AR	12.8%	12.8%	12.8%	1.00	87.2%	87.2%
CR	5.33%	4.9%	10.0%	2.15	95.1%	90.0%
AR and CR	8.17%	17.4%	24.2%	2.25	82.6%	75.8%

Table 19: Comparison of impact of reference resolution techniques on link level (ACE5)

	Link change rate (AR) and link reduction rate (CR, AR & CR)	Entities impacted by routine			Entities not impacted by routine (node weight = 1)	
		Amount	Total node weight carried	Average node weight	Amount	Total node weight carried
Genre	**AR (relaxed definition)**					
Newswire	9.6%	9.6%	9.6%	1	90.4%	90.4%
Broadcast n.	23.7%	23.7%	23.7%	1	76.3%	76.3%
Broadcast con.	20.4%	20.4%	20.4%	1	79.6%	79.6%
Telephone	53.5%	53.5%	53.5%	1	46.5%	46.5%
Usenet	39.7%	39.7%	39.7%	1	60.3%	60.3%
Weblogs	22.9%	22.9%	22.9%	1	77.1%	77.1%
Average	28.3%	28.3%	28.3%	1	71.7%	71.7%
	CR (Names and Nominals)					
Newswire	8.2%	7.5%	17.4%	2.33	92.5%	82.6%
Broadcast n.	4.2%	5.8%	12.2%	2.11	94.2%	87.8%
Broadcast con.	8.3%	9.2%	20.7%	2.26	90.8%	79.3%
Telephone	2.1%	6.9%	14.3%	2.07	93.1%	85.7%
Usenet	5.6%	7.8%	19.1%	2.45	92.2%	80.9%
Weblogs	4.6%	4.6%	11.1%	2.40	95.4%	88.9%
Average	5.5%	7.0%	15.8%	2.27	93.0%	84.2%
	AR + CR (incl. CR on anaphora)					
Newswire	10.9%	8.0%	18.9%	2.36	92.0%	81.1%
Broadcast n.	9.7%	7.8%	17.5%	2.24	92.2%	82.5%
Broadcast con.	13.4%	10.3%	23.7%	2.30	89.7%	76.3%
Telephone	19.0%	14.3%	33.4%	2.33	85.7%	66.6%
Usenet	18.4%	10.3%	28.7%	2.79	89.7%	71.3%
Weblogs	8.6%	6.0%	14.6%	2.44	94.0%	85.4%
Average	13.3%	9.5%	22.8%	2.41	90.6%	77.2%

2.7.1.3 *Impact of Reference Resolution on Network Data and Network Analysis Results*

In the ground truth data for this project, the information about entities and relations is provided as unambiguous, numerical identifiers. This situation is representative for working with social network data where each truly distinct node has a unique key identifier, even if the identifier is anonymized. Such data are typically obtained when collecting network data via surveys and participating observations. However, for semantic network data, unique node identifiers are often not available. In these situations, node names (token surface form) are often used as identifiers. Consequently, nodes matching in spelling are considered as identical nodes. For practical applications this means that when the network analysis tool encounters a node with the exact same spelling as a previously registered node, the software does not add another node to its data registry, but increases the weight of the previously found node accordingly. This is common procedure in many SNA tools and libraries. For example, when extracting network data with AutoMap, nodes are aggregated based on their spelling; regardless of capitalization. We have used this approach in a prior study on the impact of reference resolution on network data (Diesner & Carley, 2009a). This approach returns correct results if all instances of an entity are consistently referred to by the same name, and this name does not coincide with the name of a different entity. Problems with this approach occur in the cases of homographs and heteronyms (same spelling, different meaning), which cannot be disambiguated based on orthography. For example, if the term "she" is found in multiple files and cannot be resolved or disambiguated, all instances of this node are collected in one node labeled "she". For this project, I deviate from this common procedure in order to isolate the impact of RR on network data analysis while excluding the impact of coincidentally matching spellings of actually distinct nodes. This strict definition of node uniqueness is realized by using the entity mention IDs provided in ACE as node identifiers, and the heads of these entities as node names. In order to identify how disambiguating different entities with the same spelling matters for network analysis, I am also providing an empirical comparison of both approaches to determining node uniqueness (node identity based on ID versus node identity based on spelling).

In order to analyze the impact of AR and CR on network data and respective analysis results, I created one network per genre and one for the entire corpus after applying the reference resolution techniques individually and combined. These tests are conducted for ACE5 only. The networks are directed, weighted graphs. I used the ORA software to compute a selected set of frequently used network analysis measures on these graphs. These metrics are defined in Table 154. Since some of these metrics are only defined for symmetric, binary graphs, ORA internally converts the input data accordingly.

Network analysis is particularly sensitive to the connectivity and weight of nodes and links. These two characteristics impact a node's prominence and importance in the graph as well as the overall network structure. In the analysis on the link level, nodes were only embedded in dyads (regular links), whereas on the network level, a node can be linked to multiple other unique nodes, and the node degree (number of direct links per node) will increase accordingly. For the analysis on the entity and link level, the impact of heavy "outliers" (hubs) can be diluted by computing averaged degree values, while on the network level, nodes with a high degree have a strong impact on the overall network (Barabási & Albert, 1999).

Table 20 to Table 26 show the network analysis results in dependence of the RR techniques. The last three columns in each of these tables show the change from the raw data to AR, to CR, and to AR plus CR. For resolving anaphora on the network level, I used the full set of entities processed with AR. Therefore, it is possible that pronouns get resolved by nodes that were not yet present in the network such that the number of unique nodes in the network can increase from the raw data to data after AR. The following trends are observed for all genres (Table 20 to Table 25) and the entire network (all genres, Table 26): the number of nodes, links and components (strong and weak) decreases when applying each and both RR routines. Using the RR techniques leads to an increase in density, degree centralization, connectedness, transitivity, global efficiency, clustering coefficients, average distance and diffusion. All of these increases and decreases are stronger after applying CR than after applying AR (the opposite is true only for telephone data), and stronger for using AR plus CR than for using CR only. Efficiency and fragmentation are only marginally impacted and only if AR and CR are both applied. The outcomes for network levels, eigenvector centralization and average speed show changes per genre, but no clear trends in terms of increase or decrease.

The betweenness centralization of all networks was zero, which I assume to be due to the sparseness of the data. This assumption is supported by the fact that density values are consistently low. Also, closeness centralization was zero except for one genre. The network diameter equaled the number of nodes in all cases. Therefore, the three abovementioned network centralization metrics as well as the diameter are not presented in the result tables. The eigenvector centralization could not be computed on some of these networks in ORA, and is not reported if not available.

Table 20: Impact of reference resolution techniques on network properties, newswire data

Measure	Raw	AR	CR	AR & CR	Raw to AR	Raw to CR	Raw to AR & CR
Link Count	2,669	2,667	2,451	2,390	0%	-8%	-10%
Node Count	4,596	4,447	2,994	2,770	-3%	-35%	-40%
Component Count Strong	4,596	4,447	2,986	2,760	-3%	-35%	-40%
Component Count Weak	1,937	1,795	638	512	-7%	-67%	-74%
Network Levels	4	5	6	6	25%	50%	50%
Density	0.0001	0.0001	0.0003	0.0003	0%	200%	200%
Network Centr. Degree	0.0001	0.0003	0.0009	0.0031	200%	800%	3000%
Network Centr. Eigenvector	1.00	1.00	0.89	0.80	0%	-11%	-20%
Density Clustering Coeff.	0.001	0.002	0.005	0.011	64%	391%	918%
Average Distance	1.13	1.14	1.62	1.66	1%	44%	47%
Average Speed	0.89	0.88	0.62	0.60	-1%	-30%	-32%
Transitivity	0.02	0.02	0.02	0.04	45%	24%	146%
Diffusion	0.0001	0.0002	0.0005	0.0006	100%	400%	500%
Fragmentation	1.000	1.000	0.995	0.994	0%	0%	-1%
Connectedness	0.000	0.000	0.005	0.006	0%	1075%	1450%
Efficiency Global	0.0003	0.0003	0.0018	0.0023	0%	500%	667%
Efficiency	0.991	0.991	0.995	0.994	0%	0%	0%
Hierarchy	1.000	1.000	0.997	0.996	0%	0%	0%

Upper Boundedness	0.69	0.67	0.18	0.20	-3%	-74%	-72%
Interdependence	0	0.0001	0.0002	0.0002	-	-	-

Table 21: Impact of reference resolution techniques on network properties, broadcast news data

Measure	Raw	AR	CR	AR & CR	Raw to AR	Raw to CR	Raw to AR & CR
Link Count	2,008	1,999	1,925	1,821	0%	-4%	-9%
Node Count	3,576	3,285	2,920	2,519	-8%	-18%	-30%
Component Count Strong	3,576	3,283	2,920	2,519	-8%	-18%	-30%
Component Count Weak	1,572	1,295	1,015	753	-18%	-35%	-52%
Network Levels	4	5	4	4	25%	0%	0%
Density	0.0002	0.0002	0.0002	0.0003	0%	0%	50%
Network Centr. Degree	0.0003	0.0006	0.0007	0.0021	100%	133%	600%
Network Centr. Eigenvector	0.97	0.96	0.98	0.74	-2%	1%	-24%
Density Clustering Coeff.	0.000	0.001	0.002	0.010	-	-	-
Average Distance	1.10	1.16	1.24	1.26	5%	12%	15%
Average Speed	0.91	0.86	0.81	0.79	-5%	-11%	-13%
Transitivity	0.00	0.01	0.02	0.08	-	-	-
Diffusion	0.0002	0.0002	0.0003	0.0004	0%	50%	100%
Fragmentation	1.000	0.999	0.999	0.998	0%	0%	0%
Connectedness	0.000	0.001	0.001	0.002	50%	175%	300%
Efficiency Global	0.0004	0.0005	0.0007	0.001	25%	75%	150%
Efficiency	0.993	0.995	0.993	0.984	0%	0%	-1%
Hierarchy	1.000	0.999	1.000	1.000	0%	0%	0%
Upper Boundedness	0.73	0.76	0.37	0.47	4%	-50%	-35%
Interdependence	0.0001	0.0001	0.0002	0.0002	0%	100%	100%

Table 22: Impact of reference resolution techniques on network properties, broadcast conversations data

Measure	Raw	AR	CR	AR & CR	Raw to AR	Raw to CR	Raw to AR & CR
Link Count	1,656	1,650	1,520	1,438	0%	-8%	-13%
Node Count	2,872	2,648	2,077	1,776	-8%	-28%	-38%
Component Count Strong	2,871	2,646	2,075	1,774	-8%	-28%	-38%
Component Count Weak	1,220	1,006	589	404	-18%	-52%	-67%
Network Levels	4	4	5	5	0%	25%	25%
Density	0.0002	0.0002	0.0004	0.0005	0%	100%	150%
Network Centr. Degree	0.0002	0.0006	0.001	0.0032	200%	400%	1500%
Network Centr. Eigenvector	0.97	0.96	0.76	0.92	-1%	-21%	-4%
Density Clustering Coeff.	0.000	0.001	0.003	0.011	100%	750%	2725%
Average Distance	1.11	1.15	1.34	1.36	4%	21%	23%
Average Speed	0.90	0.87	0.75	0.73	-4%	-17%	-19%
Transitivity	0.01	0.01	0.02	0.06	46%	266%	852%
Diffusion	0.0002	0.0003	0.0005	0.0006	50%	150%	200%
Fragmentation	1.000	0.999	0.997	0.994	0%	0%	-1%
Connectedness	0.001	0.001	0.004	0.006	80%	600%	1060%

Efficiency Global	0.0005	0.0006	0.0016	0.0024	20%	220%	380%
Efficiency	0.995	0.996	0.995	0.992	0%	0%	0%
Hierarchy	1.000	0.999	0.999	0.999	0%	0%	0%
Upper Boundedness	0.76	0.69	0.20	0.22	-9%	-74%	-72%
Interdependence	0.0001	0.0001	0.0003	0.0003	0%	200%	200%

Table 23: Impact of reference resolution techniques on network properties, telephone conversations data

Measure	Raw	AR	CR	AR & CR	Raw to AR	Raw to CR	Raw to AR & CR
Link Count	746	739	730	604	-1%	-2%	-19%
Node Count	1,377	1,079	1,161	799	-22%	-16%	-42%
Component Count Strong	1,377	1,077	1,161	797	-22%	-16%	-42%
Component Count Weak	631	347	435	212	-45%	-31%	-66%
Network Levels	4	4	4	4	0%	0%	0%
Density	0.0004	0.0006	0.0005	0.0009	50%	25%	125%
Network Centr. Degree	0.0011	0.0048	0.002	0.0072	336%	82%	555%
Network Centr. Eigenvector	0.9993	0.9813	0.7053	0.9562	-2%	-29%	-4%
Density Clustering Coeff.	0.000	0.003	0.000	0.009	-	-	-
Average Distance	1.08	1.24	1.13	1.27	15%	5%	17%
Average Speed	0.93	0.80	0.88	0.79	-13%	-5%	-15%
Transitivity	0.00	0.02	0.00	0.07	-	-	-
Diffusion	0.0004	0.0008	0.0006	0.0012	100%	50%	200%
Fragmentation	0.999	0.995	0.998	0.992	0%	0%	-1%
Connectedness	0.001	0.005	0.002	0.008	456%	122%	778%
Efficiency Global	0.0009	0.0027	0.0015	0.0041	200%	67%	356%
Efficiency	1.000	0.997	0.994	0.991	0%	-1%	-1%
Hierarchy	1.000	0.997	1.000	0.996	0%	0%	0%
Upper Boundedness	0.76	0.76	0.30	0.54	0%	-60%	-29%
Interdependence	0.0001	0.0002	0.0005	0.0005	100%	400%	400%

Table 24: Impact of reference resolution techniques on network properties, usenet data

Measure	Raw	AR	CR	AR & CR	Raw to AR	Raw to CR	Raw to AR & CR
Link Count	858	846	811	705	-1%	-5%	-18%
Node Count	1,547	1,322	1,208	936	-15%	-22%	-39%
Component Count Strong	1,547	1,322	1,208	936	-15%	-22%	-39%
Component Count Weak	692	479	402	247	-31%	-42%	-64%
Network Levels	3	6	4	4	100%	33%	33%
Density	0.0004	0.0005	0.0006	0.0008	25%	50%	100%
Network Centr. Degree	0.0008	0.0016	0.0022	0.0067	100%	175%	738%
Network Centr. Eigenvector	1.00	0.98	0.99	0.98	-2%	-1%	-2%
Density Clustering Coeff.	0.002	0.002	0.003	0.011	0%	53%	453%
Average Distance	1.08	1.25	1.24	1.33	16%	16%	24%
Average Speed	0.93	0.80	0.80	0.75	-14%	-14%	-19%

Transitivity	0.03	0.01	0.02	0.05	-62%	-38%	38%
Diffusion	0.0004	0.0006	0.0007	0.0011	50%	75%	175%
Fragmentation	0.999	0.997	0.997	0.993	0%	0%	-1%
Connectedness	0.001	0.003	0.003	0.007	211%	222%	667%
Efficiency Global	0.0008	0.0017	0.0018	0.0036	113%	125%	350%
Efficiency	0.985	0.998	0.996	0.993	1%	1%	1%
Hierarchy	1.000	1.000	1.000	1.000	0%	0%	0%
Upper Boundedness	0.68	0.85	0.29	0.47	25%	-57%	-31%
Interdependence	0.0001	0.0002	0.0005	0.0005	100%	400%	400%

Table 25: Impact of reference resolution techniques on network properties, blog data

Measure	Raw	AR	CR	AR & CR	Raw to AR	Raw to CR	Raw to AR & CR
Link Count	766	766	732	703	0%	-4%	-8%
Node Count	1,407	1,331	1,137	1,031	-5%	-19%	-27%
Component Count Strong	1,407	1,331	1,137	1,031	-5%	-19%	-27%
Component Count Weak	643	567	412	340	-12%	-36%	-47%
Network Levels	3	4	4	4	33%	33%	33%
Density	0.0004	0.0004	0.0006	0.0007	0%	50%	75%
Network Centr. Degree	0.0003	0.0009	0.0015	0.0052	200%	400%	1633%
Network Centr. Eigenvector	0.79	0.98	0.94	0.95	25%	20%	21%
Density Clustering Coeff.	0.001	0.001	0.004	0.009	0%	236%	755%
Average Distance	1.06	1.10	1.20	1.24	4%	13%	17%
Average Speed	0.94	0.91	0.83	0.80	-4%	-12%	-15%
Transitivity	0.02	0.02	0.03	0.06	-35%	19%	144%
Diffusion	0.0004	0.0005	0.0007	0.0008	25%	75%	100%
Fragmentation	0.999	0.999	0.997	0.997	0%	0%	0%
Connectedness	0.001	0.001	0.003	0.003	44%	200%	278%
Efficiency Global	0.0008	0.001	0.0017	0.0022	25%	113%	175%
Efficiency	0.987	0.994	0.993	0.989	1%	1%	0%
Hierarchy	1.000	1.000	1.000	1.000	0%	0%	0%
Upper Boundedness	0.71	0.78	0.37	0.50	10%	-47%	-30%
Interdependence	0.0001	0.0001	0.0005	0.0005	0%	400%	400%

The results for disambiguating and consolidating nodes based on node IDs versus node spelling differ strongly (Table 26): with the spelling based approach, for 2/3 of the considered measures, AR and CR exhibit opposite effects with respect to increasing or decreasing the value of a network metric, AR causes a greater change rate than CR, and the joint impact of AR and CR is moderate in most cases (for 13 of 20 measures, the combined change rate is 10% and less). These effects are consistent with our previous findings (Diesner & Carley, 2009a), but differ starkly from disambiguating nodes based on their actual ID: there, AR and CR both either increase or decrease a metric (except for upper boundedness), CR has a stronger impact than AR, and the joint impact of AR and CR is much larger than with the alternative approach (13 out of 20 measures have a change rate of 10% and more, 5 metrics have a change rate of more than 100%). In summary, the results for node disambiguation approaches suggest that consolidating nodes based on spelling leads to network data, analysis results and interpretations that strongly deviate from what is suggested by using ground truth data and allows for a smaller overall effect of RR.

Table 26: Impact of reference resolution techniques on network properties, node identity based on spelling versus node ID, all genres

Measure	Raw	AR	CR	AR & CR	Raw to AR	Raw to CR	Raw to AR & CR
Entire network, node disambiguation and consolidation based on node ID							
Link Count	8,703	8,667	8,169	7,661	0%	-6%	-12%
Count Node	15,375	14,112	11,497	9,831	-8%	-25%	-36%
Component Count Strong	15,374	14,106	11,487	9,817	-8%	-25%	-36%
Component Count Weak	6,695	5,489	3,491	2,468	-18%	-48%	-63%
Network Levels	4	6	6	6	50%	50%	50%
Density	0	0	0.0001	0.0001	-	-	-
Network Centr. Degree	0.0001	0.0001	0.0002	0.0009	0%	100%	800%
Network Centr., Between.	0	0	0	0	-	-	-
Density Clustering Coeff.	0.001	0.001	0.003	0.011	100%	357%	1400%
Average Distance	1.10	1.16	1.39	1.44	6%	26%	30%
Speed Average	0.91	0.86	0.72	0.70	-5%	-21%	-23%
Transitivity	0.01	0.02	0.02	0.05	41%	81%	370%
Diffusion	0	0.0001	0.0001	0.0001	-	-	-
Fragmentation	1.00	1.00	1.00	1.00	0%	0%	0%
Connectedness	0.0001	0.0002	0.0006	0.0009	100%	500%	800%
Efficiency Global	0.0001	0.0001	0.0003	0.0004	0%	200%	300%
Efficiency	0.992	0.995	0.995	0.992	0%	0%	0%
Hierarchy	1.000	0.999	0.998	0.998	0%	0%	0%
Upper Boundedness	0.72	0.75	0.22	0.27	4%	-70%	-63%
Interdependence	0	0	0.0001	0.0001	-	-	-
Entire network, node disambiguation and consolidation based on node spelling							
Link Count	6,475	6,669	6,561	6,514	3%	1%	1%
Count Node	3,299	3,518	3,215	3,323	7%	-3%	1%
Component Count Strong	2,780	2,988	2,638	2,763	7%	-5%	-1%
Component Count Weak	165	170	124	130	3%	-25%	-21%
Network Levels	21	21	20	23	0%	-5%	10%
Density	0.0006	0.0005	0.0006	0.0006	-17%	0%	0%
Network Centr. Degree	0.0009	0.0008	0.0011	0.0008	-11%	22%	-11%
Network Centr., Between.	0.029	0.038	0.033	0.037	33%	14%	30%
Density Clustering Coeff.	0.013	0.019	0.028	0.045	47%	110%	240%
Average Distance	5.69	6.31	5.80	6.47	11%	2%	14%
Speed Average	0.18	0.16	0.17	0.15	-10%	-2%	-12%
Transitivity	0.04	0.04	0.05	0.04	-9%	8%	3%

Diffusion	0.1891	0.1719	0.2160	0.1905	-9%	14%	1%
Fragmentation	0.21	0.21	0.16	0.17	0%	-22%	-20%
Connectedness	0.7931	0.7926	0.8391	0.8342	0%	6%	5%
Efficiency Global	0.1873	0.1757	0.1993	0.1833	-6%	6%	-2%
Efficiency	0.999	0.999	0.999	0.999	0%	0%	0%
Hierarchy	0.931	0.930	0.919	0.921	0%	-1%	-1%
Upper Boundedness	0.64	0.58	0.67	0.60	-10%	5%	-6%
Interdependence	0.0001	0.0001	0.0001	0.0001	0%	0%	0%

For practical applications of network analysis, in addition to computing network level metrics, people are often interested in identifying the set of nodes that score highest on a certain measure or a set of measures. This procedure is also called “key player analysis”. I perform key player analyses on the data by using ORA to compute several network analytical measures for every node per network, and comparing the top five ranking nodes after each and both RR technique were applied (Table 27, tying nodes listed in alphabetical order).

For RR based on actual node IDs, the results show that the set of key entities identified when not applying any RR technique are completely different from the key entities found after applying RR (Table 27). When performing both, AR and CR, the key entities for betweennees centrality and in-degree centrality are similar to the key entities found after using CR only, and the key players with respect to inverse closeness centrality and out-degree centrality resemble those identified by using AR only. Since the values per measure and node are overall higher and more often different from zero for betweennees centrality and in-degree centrality than for inverse closeness centrality and out-degree centrality, the findings for similarities between CR and AR plus CR are more robust than the similarities after using AR. For practical applications, this means that performing at least CR will cause a major change in the network data, which resembles the ground truth more closely than using no RR or AR only.

Several two top scoring nodes in the raw data are pronouns, e.g. “which”, “she”, “all”, and “they”, which are unlikely to present the actual agents who drive the dynamics of a system. Ironically, the top scoring node w.r.t. out-degree centrality is “we”. What looks like a mistake represents the fact that especially in the accounts of spoken language as well as in the social media data data, “we” is a frequently occurring entity that sometimes cannot be resolved via AR, but consolidated via CR.

When consolidating nodes based on spelling, the set of key players identified with and without using any RR techniques are highly similar to each other. Interpreting this finding together with the outcome of the network level analyses suggests the when normalizing nodes based on spelling only, RR makes a much smaller difference with respect to changes in network metrics and identified key players than when normalizing nodes based on actual node IDs. Taking this interpretation a step further implies that if only key players and a certain set of measures (listed at end of this sentence) are to be computed, conducting any RR technique is not worthwhile if nodes are normalized based on spelling (number of nodes, number of links, strong components, network levels, density, transitivity, diffusion, connectedness, global efficiency, efficiency, hierarchy, upper boundedness, interdependence). However, the results obtained this way do not resemble findings based on ground truth data, i.e. nodes disambiguated based on node IDs.

Table 27: Key entities, node identity based on spelling versus node ID, all genres, ACE5

	Node disambiguation and consolidation based on node ID				**Node disambiguation and consolidation based on node spelling**			
Rank	Betweenness centrality	Inverse closeness centrality	In-degree centrality	Out-degree centrality	Betweenness centrality	Inverse closeness centrality	In-degree centrality	Out-degree centrality
	Raw							
1	home	soldiers	Washington	all	Iraq	director	U.S	his
2	Byrds Creek	she	area	ambassadors	I	founder	Iraqi	forces
3	base	boy	home	Protesters	they	chairman	Iraq	troops
4	streets	forces	which	diplomats	his	Chiefs of Staff	Baghdad	my
5	mosque	forces	Tuesday	Iraqis	area	Giuliani	there	I
	AR							
1	Judy	parents	company	Judy	Iraq	Roger	U.S	forces
2	Ringo Langly	Judy	headquarters	GF	troops	guy	Iraqi	troops
3	GF	Annie J.S.	base	dogbirdh@…	forces	executive	Iraq	people
4	kramer	guy	group	Britt	family	director	Baghdad	officials
5	dad	Britt	US	Annie J.S.	people	chairman	city	President
	CR							
1	Indonesia	forces	country	Stig Toefting	his	director	U.S	his
2	Iraq	Buildings	Palestinian	terrorist	people	Council	Iraq	forces
3	Iraqi	source	Iraqi	bomber	Iraq	head	Iraqi	troops
4	city	TV2	American	Iraq	I	Protesters	Baghdad	my
5	Stig Toefting	Copenhagen	Indonesia	troops	Baghdad	Task Force	country	I
	AR & CR							
1	Indonesia	parents	country	we	Iraq	Council	U.S	troops
2	Iraqi	Judy	Palestinian	private	people	head	Iraq	forces
3	Iraqi	mother	Indonesia	Marwan B.	President	Shaq	Iraqi	people
4	Stig Toefting	Mildred	Iraqi	Judy	U.S	Copenhagen	Baghdad	officials
5	city	industry	U.S	GF	troops	TV2	country	President

2.7.1.4 Simulation of impact of reference resolution error rates

The last research question for the RR project is about the impact of changes in the accuracy rates of AR and CR on network data. I use the following procedure to study the effect of introducing typical RR errors into ground truth data: my review of typical error rates of current, publically available and top performing RR tools has shown that precision is about ten percent higher than recall, and that recall and precision range between 55% to 85%, and 65% to 95%, respectively (Table 4). Based on this review of empirical results, I defined the following four levels of accuracy rates as shown in Table 28 for experimentation. Next, I assume that the ground truth data are the gold standard against which the performance of a reference resolution tool would be compared in order to assess the tool's accuracy. This procedure resembles the way accuracy assessment is done in NLP. Based on this assumption, I introduce errors into the ground truth data such that the resulting data have the error rates specified in Table 28 as follows: I generate false negatives by removing randomly selected links from the ground truth until a given recall rate has been reached. Once this is done, I add false positives into the data by connecting nodes that are not linked in the ground truth data, but are defined as valid nodes in the ground truth. The weight of added links is selected proportionally to the distribution of link weights in the ground truth, which differs per RR technique and was also handled this way. Once the data with the given error rates were constructed, I performed the same network analysis on them as presented in the previous section in order to allow for comparability and generalizability of the findings. These analyses were performed for the ACE5 data on the entire corpus level.

Table 28: Accuracy rates for reference resolution for experiments

	Precision	Recall	F
Accuracy I	55	65	60
Accuracy II	65	75	70
Accuracy III	75	85	80
Accuracy VI	85	95	90

Table 29 to Table 31 show the previously used network metrics in dependence of the increase in accuracy by 10% for the first four columns, and the difference between the values computed on the ground truth data to each accuracy setting in the last four columns. The following trends can be observed for AR, CR and AR plus CR: The most common effect is that increases in accuracy lead to decreases in the underestimation of the following metrics (listed by decreasing amount of underestimation): upper boundedness, transitivity, clustering coefficient, the number of strong and weak components, the number of nodes and links, and average speed. For either and both RR techniques, increases in accuracy also lead to decreases in the overestimates of the following metrics (listed by decreasing amount of overestimating): connectedness, diffusion, global efficiency, network levels, and degree centralization. Improving the accuracy for each and both RR techniques has virtually no impact of network density, fragmentation and efficiency.

The results show that overall, even small error rates can cause huge changes in the value of network metrics in comparison to ground truth data, which herein is assumed to represent 100% correct RR. To illustrate this effect, I have underlined the conditions under which changes occur and where the difference between the true value and the value obtained using a certain error rate is equal to or less than 10%. This applies only to metrics which did show no clear trend in how they change depending on RR techniques as discussed in section 2.7.1.3, namely efficiency, fragmentation, network levels, and speed, or requires the highest accuracy rate tested to achieve this effect, which applies to diffusion and the number of links only.

Table 29: Change in network properties depending on error rates for AR

Measure	Accu-racy I	Accu-racy II	Accu-racy III	Accu-racy IV	Ground Truth	Acc I to GT	Acc II to GT	Acc III to GT	Acc IV to GT
Connectedness	0.0034	0.0040	0.0005	0.0003	0.0002	1600%	1900%	150%	50%
Efficiency Global	0.0006	0.0005	0.0002	0.0002	0.0001	500%	400%	100%	100%
Diffusion	0.0001	0.0001	0.0001	0.0001	0.0001	0%	0%	0%	0%
Network Levels	10	9	8	6	6	67%	50%	33%	0%
Nw. Centr. Degree	0.0003	0.0002	0.0002	0.0002	0.0001	200%	100%	100%	100%
Upper Boundedness	0.11	0.06	0.44	0.60	0.75	-86%	-92%	-41%	-19%
Transitivity	0.001	0.002	0.003	0.010	0.016	-92%	-89%	-81%	-39%
Average Distance	1.90	1.76	1.52	1.27	1.16	63%	51%	30%	9%
Density Clus. Coeff.	0.0004	0.0005	0.0006	0.0013	0.0014	-71%	-64%	-57%	-7%
Comp. Count Weak	2,613	3,110	3,775	4,654	5,489	-52%	-43%	-31%	-15%
Average Speed	0.53	0.57	0.66	0.78	0.86	-39%	-34%	-23%	-9%
Node Count	9,973	10,642	11,422	12,387	14,112	-29%	-25%	-19%	-12%
Comp. Count Strong	9,971	10,640	11,419	12,383	14,106	-29%	-25%	-19%	-12%
Link Count	7,368	7,539	7,662	7,765	8,667	-15%	-13%	-12%	-10%
Fragmentation	0.997	0.996	1.000	1.000	1.000	0%	0%	0%	0%
Efficiency	1.000	1.000	1.000	0.998	0.995	0%	0%	0%	0%
Hierarchy	1.00	1.00	1.00	1.00	1.00	0%	0%	0%	0%
Density	0.0001	0.0001	0.0001	0.0001	0	-	-	-	-

Table 30: Change in network properties depending on error rates for CR

Measure	Accu-racy I	Accu-racy II	Accu-racy III	Accu-racy IV	Ground Truth	Acc I to GT	Acc II to GT	Acc III to GT	Acc IV to GT
Connectedness	0.2014	0.1277	0.0416	0.0013	0.0006	>33tsd%	>21tsd%	6833%	117%
Efficiency Global	0.0122	0.0075	0.0024	0.0004	0.0003	3967%	2400%	700%	33%
Diffusion	0.0003	0.0002	0.0002	0.0001	0.0001	200%	100%	100%	0%
Network Levels	15	11	11	8	6	150%	83%	83%	33%
Nw. Centr. Degree	0.0003	0.0005	0.0004	0.0003	0.0002	50%	150%	100%	50%
Upper Boundedness	0.00	0.00	0.01	0.13	0.22	-98%	-98%	-96%	-38%
Transitivity	0.001	0.004	0.007	0.012	0.021	-95%	-81%	-68%	-40%
Average Distance	2.99	2.33	2.04	1.56	1.39	115%	68%	47%	12%
Density Clus. Coeff.	0.0004	0.0008	0.0018	0.0020	0.0032	-88%	-75%	-44%	-38%
Comp. Count Weak	1,558	1,914	2,387	2,965	3,491	-55%	-45%	-32%	-15%
Average Speed	0.33	0.43	0.49	0.64	0.72	-54%	-40%	-32%	-11%
Node Count	8,421	8,924	9,556	10,195	11,497	-27%	-22%	-17%	-11%
Comp. Count Strong	8,416	8,922	9,549	10,191	11,487	-27%	-22%	-17%	-11%
Link Count	6,968	7,100	7,236	7,322	8,169	-15%	-13%	-11%	-10%
Fragmentation	0.799	0.872	0.958	0.999	0.999	-20%	-13%	-4%	0%

Efficiency	1.000	1.000	1.000	0.998	0.995	1%	1%	1%	0%
Hierarchy	1.00	1.00	1.00	1.00	1.00	0%	0%	0%	0%
Density	0.0001	0.0001	0.0001	0.0001	0.0001	0%	0%	0%	0%

Table 31: Change in network properties depending on error rates for AR and CR

Measure	Accu-racy I	Accu-racy II	Accu-racy III	Accu-racy IV	Ground Truth	Acc I to GT	Acc II to GT	Acc III to GT	Acc IV to GT
Connectedness	0.3318	0.2704	0.1608	0.0046	0.0009	>36tsd%	29tsd%	>17tsd%	411%
Efficiency Global	0.0225	0.0191	0.0095	0.0008	0.0004	5525%	4675%	2275%	100%
Diffusion	0.0004	0.0004	0.0002	0.0001	0.0001	300%	300%	100%	0%
Network Levels	18	15	16	9	6	200%	150%	167%	50%
Nw. Centr. Degree	0.0012	0.0009	0.001	0.001	0.0009	33%	0%	11%	11%
Upper Boundedness	0.00	0.01	0.00	0.07	0.27	-98%	-98%	-98%	-74%
Transitivity	0.007	0.008	0.018	0.027	0.053	-87%	-85%	-65%	-49%
Average Distance	3.14	3.13	2.37	1.69	1.44	118%	117%	65%	17%
Density Clus. Coeff.	0.0026	0.0027	0.0051	0.0060	0.0105	-75%	-74%	-51%	-43%
Comp. Count Weak	1,088	1,285	1,642	2,114	2,468	-56%	-48%	-33%	-14%
Average Speed	0.32	0.32	0.42	0.59	0.70	-54%	-54%	-39%	-15%
Node Count	7,394	7,785	8,268	8,819	9,831	-25%	-21%	-16%	-10%
Comp. Count Strong	7,394	7,780	8,265	8,812	9,817	-25%	-21%	-16%	-10%
Link Count	6,509	6,723	6,800	6,866	7,661	-15%	-12%	-11%	-10%
Fragmentation	0.668	0.730	0.839	0.995	0.999	-33%	-27%	-16%	0%
Efficiency	1.000	1.000	1.000	0.999	0.992	1%	1%	1%	1%
Hierarchy	1.00	1.00	1.00	1.00	1.00	0%	0%	0%	0%
Density	0.0001	0.0001	0.0001	0.0001	0.0001	0%	0%	0%	0%

In order to test the qualitative impacts of the given error rates, I performed the same type of key player analysis as presented earlier in this chapter. The outcomes ((Table 32 to Table 34) differ from what the quantitative analysis on the simulation of the impact of RR error rates had suggested: for both RR techniques, individually and combined, there is a large amount of overlap in key entities between the ground truth data and key entities found at lower RR accuracy rates, especially with respect to node degree centrality. This is even true for fairy low accuracy rates. This finding suggests that the set of key players is less sensitive towards changes in accuracy rates than network analytical measures. One possible explanation for this finding is the very nature of highly central nodes: their ratio among all nodes is very low such that dropping them in a randomized or based on node degree fashion is very low. However, since their node degree is exponentially higher than the degree of non-central nodes, removing a fraction of the links from central nodes or nodes around them will have only a minor impact on these nodes, even though computing network metrics on the modified graph is more sensitive to these modification as shown in the previous section. Finally, the key players are similar for CR and AR plus CR, but rather different set of key players is identified when performing AR only. This suggests that AR has a smaller impact on the combined results than CR does.

Table 32: Change in key players depending on error rates for AR

Betweenness centrality	Inverse closeness centrality	In-degree centrality	Out-degree centrality
Accuracy I			
Judy	organization	Judy	Annie Juhlyn Simon
dogbirdh...@yahoo.com	Lynn	company	Judy
base	Jabaliya	streets	dogbirdh...@yahoo.com
Annie Juhlyn Simon	area	U.S	Barbara Sz.
GF	Universal Orlando	headquarters	roommate
Accuracy II			
Ringo Langly	industry	group	dogbirdh...@yahoo.com
roommate	grandmother	BIL	Britt
base	Giuliani	fort hood	GF
nephew	Rudolph Giuliani	Washington DC	Mark
man	companion	headquarters	Judy
Accuracy III			
teacher	possessions	base	Judy
Judy	body	fort hood	GF
Mildred	guy	company	dogbirdh...@yahoo.com
dogbirdh...@yahoo.com	closet	US	Annie Juhlyn Simon
students	parents	group	Britt
Accuracy VI			
Judy	head	headquarters	Judy
teacher	court	company	GF
AIG	parents	group	dogbirdh...@yahoo.com
tracy	Judy	Washington DC	Annie Juhlyn Simon
court	Annie Juhlyn Simon	fort hood	Barbara Sz.
Ground truth			
Judy	parents	company	Judy
Ringo Langly	Judy	headquarters	GF
GF	Annie Juhlyn Simon	base	dogbirdh@ yahoo.com
kramer	guy	group	Britt
dad	Britt	US	Annie Juhlyn Simon

Table 33: Change in key players depending on error rates for CR

Betweenness centrality	Inverse closeness centrality	In-degree centrality	Out-degree centrality
Accuracy I			
Agartala	we	Palestinian	Stig Toefting
son	who	Indonesia	soldiers
Indonesia	forces	people	bomber
people	new york	Israeli	members
members	reserves	US	Vivendi Universal
Accuracy II			
American	troops	country	Giuliani
Iraqi	rats	Palestinian	terrorist
city	Diller	American	you
Iraqi	resistance	Iraqi	McCarthy
Patriot	McCarthy	Indonesia	Iraq
Accuracy III			

Stig Toefting	neighborhood	country	Stig Toefting
Iraq	North Korean	US	members
Israel	Stig Toefting	Palestinian	terrorist
crossing	parliament	American	Iraq
Denmark	ambassador	American	North Korean
Accuracy VI			
American	its	Iraqi	Giuliani
Indonesia	park	American	Iraq
baby	Vivendi Universal	country	Indonesia
Iraqi	officials	people	michael sears
williams	troops	Palestinian	terrorist
Ground truth			
Indonesia	forces	country	Stig Toefting
Iraq	Buildings	Palestinian	terrorist
Iraqi	source	Iraqi	bomber
city	TV2	American	Iraq
Stig Toefting	Copenhagen	Indonesia	troops

Table 34: Change in key players depending on error rates for AR and CR

Betweenness centrality	Inverse closeness centrality	In-degree centrality	Out-degree centrality
Accuracy I			
Iraqi	ambassador	American	private
abby	your	country	girlfriend
house	Karim	American	Britt
Baghdad	minister	people	JBELLU...@COMCAST.
we	woman	Indonesia	people
Accuracy II			
mother	secretary	people	private
Security Council	troops	Iraqi	your
troop	soldiers	American	terrorist
private	state	Israel	Britt
Saudi	U.S	group	Judy
Accuracy III			
Hebron	street	country	we
American	clerics	Palestinian	Stig Toefting
prize	demonstrators	Israeli	private
Northwestern	minority	Indonesia	Britt
workers	area	Israel	terrorist
Accuracy VI			
Britt	boy	country	we
Baghdad	Mildred	US	Mildred
Indonesia	village	Indonesia	Judy
American	industry	Palestinian	Stig Toefting
court	source	American	mother
Ground truth			
Indonesia	parents	country	we
Iraqi	Judy	Palestinian	private
Iraqi	mother	Indonesia	Marwan B.

Stig Toefting	Mildred	Iraqi	Judy
city	industry	U.S	GF

2.7.1.5 Answers to research questions

The presented results for reference resolution on the entity or node, link and network data level suggest the answers to my research questions presented in Table 35. All numbers reported in this summary are averages.

Table 35: Answers to research questions 1-3

Level of analysis	**How large is the impact of the RR techniques?**	**Which routine, AR or CR, is more effective in achieving these effects?**	**Is combining AR and CR more effective than either technique alone?**
1. Entity level	Performing RR alters the identity and/or weight of 76% of all entity mentions. The entity weight is increased from 1.0 to 4.9 with AR, to 4.5 with CR, and to 5.8 with AR and CR. Less than 18% of the unique entities are impacted by RR; they carry more than 79% of the total entity weight.	CR w.r.t. the amount of entities changed. AR w.r.t. increasing the weight of impacted entities. The rate of entity reduction via CR is 45%. The rate of entity change via AR is 31%.	Yes. Combining both techniques increases the amount of entities impacted by RR by another 38%.
2. Link level	The link weight is increased from 1.0 to 2.4 by using RR. The weight of unique relations impacted by both techniques increases to less than 2.5. Less than 11% of the unique links are impacted by RR; they carry almost 23% of the total link weight.	AR. The link reduction rate due to CR is 6%. The link change rate due to AR is 23%.	Yes. When applying both techniques, 12% of all links are reduced. The impact of RR is stronger on the node level than on the link level.
3. Network level	Using RR leads to increases in network density, connectedness, transitivity, degree centralization, global efficiency, clustering coefficients, average distance and diffusion. Disambiguating nodes based on node IDs versus node spelling makes a big difference; using the latter approach leads to analysis results and interpretations that strongly deviate from the ground truth.	CR. When identifying key entities, CR closely resembles the nodes identified by using AR and CR, while applying AR only returns a completely different set of key entities.	Yes.

Question 4: How much change in network properties in due to increases in accuracy of AR and CR?

Answer 4: Even small error rates, e.g. an F value for accuracy of 90%, can cause over- and underestimations of the true network analytical values per metric of much more than 10%; often ranging up to 100% and more. In contrast to that, the identification of key entities is less sensitive towards changes in RR accuracy rates than the network analytical measures are. Also, the set of key entities is strongly impacted by CR, and less so by AR.

2.7.2 Windowing

The operationalization of "window size" for this project is the number of space separated tokens that occur between the heads of the nodes in each annotated relation. The nodes themselves are not within the window. For example, if two nodes in a link occur adjacent to each other, the window size is zero. If no head is available for an entity, which applies to all instances of the timex" class, the number of tokens between the extents of the nodes is counted.

In some ground truth data, genitive markers ('s) are separated by a single space character from the token they belong to. I use the following rule for handling this situation: These markers are disregarded from counting the size of the window. The same rule is applied to hyphens and single-character punctualization symbols, including commas.

The chosen operationalization of windowing slightly differs from another common way of measuring the length of the window, where the linked nodes are within the window. For example, if two adjacent unigrams would form a link, the window size would be two. The latter approach is used in AutoMap (Carley, Columbus, Bigrigg, & Kunkel, 2011). I chose the abovementioned operationalization in order to avoid any conflicts with entities that are multi-word expressions. Consequently, the results reported herein eliminate this source of ambiguity.

In the context of this project, the SemEval data complement the ACE datasets in several ways: first, in SemEval, different types of semantic relations are considered than in ACE. Table 36 lists the types of relations considered in SemEval along with the amount of data per type. These relations are based on prior work in semantic role labeling (Nastase & Szpakowicz, 2003). Second, in SemEval, only relations between nominals, i.e. nouns and base noun phrases, are annotated, but not between named entities or pronouns. Third, the examples in SemEval are limited to statements about real world situations. This means that negations, modalities, and opinions are exluded; all of which can be represented in ACE. Fourth, the SemEval data were collected more recently than the ACE data, and are not confined to specific genres or domains.

The drawback with this less constrained data collection approach is that one cannot know the production or release date and genre or domain of the selected texts. Finally, in ACE, the types of entities are not annotated. These differences will allow for testing the robustness of window sizes across these different aspects.

Table 36: Types of relationships and size in corpus (SemEval)

Type of Semantic Relationship	Number of Links	Ratio in Corpus
Cause-Effect	1,331	12.4%
Component-Whole	1,253	11.7%
Content-Container	732	6.8%
Entity-Destination	1,137	10.6%
Entity-Origin	974	9.1%
Instrument-Agency	660	6.2%
Member-Collection	923	8.6%
Message-Topic	895	8.4%
Other	1,864	17.4%
Product-Producer	948	8.8%

2.7.2.1 Typical window sizes and link coverage rates

The results presented in Table 37 suggest that typical window sizes as well as the ratio of links that are found when using a certain window size, which I herein refer to as coverage rate, are highly similar across the considered types of semantic relationships: for all those types, more than half of the links are found with a window size of four or less. On average, a window size of seven is needed to identify at least than 90% of the links, and with a window size of eight, over 95% of the links are retrieved. The most frequent window size that human coders apply is small, typically two or three (those values underlined in Table 37).

Table 37: Impact of type of semantic relationship on window size (SemEval)

Window Size	Per link type: Ratio of links with this size (left), Cumulative coverage of links at this size (right)											
	Cause Effect		Component Whole		Content Container		Entity Destination		Entity Origin		Instrument Agency	
0	1.4%	1.4%	12.1%	12.1%	1.2%	1.2%	0.0%	0.0%	15.8%	15.8%	4.8%	4.8%
1	11.6%	12.9%	4.5%	16.7%	7.1%	8.3%	3.8%	3.8%	0.8%	16.6%	7.1%	12.0%
2	14.0%	27.0%	40.8%	57.5%	18.7%	27.0%	26.3%	30.1%	13.0%	29.7%	19.2%	31.2%
3	20.7%	47.7%	14.1%	71.6%	32.1%	59.2%	20.4%	50.5%	18.6%	48.3%	14.2%	45.5%
4	15.3%	63.0%	8.1%	79.7%	17.1%	76.2%	22.7%	73.2%	20.4%	68.7%	8.5%	53.9%
5	10.1%	73.0%	6.7%	86.4%	11.2%	87.4%	15.8%	89.0%	13.9%	82.5%	12.0%	65.9%
6	8.9%	82.0%	5.5%	91.9%	6.0%	93.4%	5.9%	94.9%	8.4%	91.0%	10.9%	76.8%
7	7.0%	89.0%	3.3%	95.2%	1.9%	95.4%	1.8%	96.7%	3.7%	94.7%	7.3%	84.1%
8	3.5%	92.4%	1.5%	96.7%	1.9%	97.3%	1.2%	98.0%	2.2%	96.8%	4.7%	88.8%
9	2.6%	95.0%	0.9%	97.6%	1.0%	98.2%	0.6%	98.6%	1.4%	98.3%	3.2%	92.0%

10	1.6%	96.5%	1.2%	98.8%	0.1%	98.4%	0.8%	99.4%	0.4%	98.7%	2.4%	94.4%
11	0.9%	97.4%	0.6%	99.4%	0.7%	99.0%	0.4%	99.7%	0.5%	99.2%	1.8%	96.2%
12	1.1%	98.6%	0.1%	99.5%	0.1%	99.2%	0.2%	99.9%	0.1%	99.3%	1.1%	97.3%
	Member Collection		**Message Topic**		**Product Producer**		**Other**		**Average (unweighted)**			
0	2.2%	2.2%	0.7%	0.7%	12.6%	12.6%	6.8%	6.8%	5.8%	5.8%		
1	37.7%	39.9%	5.9%	6.6%	6.1%	18.7%	9.2%	16.0%	9.4%	15.1%		
2	42.7%	82.6%	22.9%	29.5%	14.9%	33.5%	21.5%	37.4%	23.4%	38.6%		
3	9.8%	92.3%	19.2%	48.7%	22.2%	55.7%	20.1%	57.5%	19.1%	57.7%		
4	3.3%	95.6%	16.1%	64.8%	16.5%	72.2%	15.0%	72.5%	14.3%	72.0%		
5	2.6%	98.2%	12.1%	76.9%	8.2%	80.4%	10.4%	82.8%	10.3%	82.3%		
6	0.8%	98.9%	7.7%	84.6%	6.1%	86.5%	6.5%	89.3%	6.7%	88.9%		
7	0.5%	99.5%	6.6%	91.2%	4.4%	90.9%	3.9%	93.2%	4.0%	93.0%		
8	0.3%	99.8%	3.1%	94.3%	2.1%	93.0%	2.1%	95.4%	2.3%	95.3%		
9	0.1%	99.9%	2.2%	96.5%	1.9%	94.9%	2.0%	97.4%	1.6%	96.8%		
10	0.0%	99.9%	1.3%	97.9%	1.3%	96.2%	0.8%	98.2%	1.0%	97.8%		
11	0.0%	99.9%	1.0%	98.9%	0.8%	97.0%	0.7%	98.9%	0.7%	98.6%		
12	0.0%	99.9%	0.6%	99.4%	0.9%	98.0%	0.4%	99.3%	0.5%	99.0%		

There are a few noteworthy differences depending on the type of semantic relationships: for "member - collection" links, which encode non-functional relationships between specific elements and some set, the window is particularly short: over 80% of nodes in a link are separated by one or two words in the text data. In contrast to that, two types of relations require a slighty larger window than the reported averages (greater by one to two words): "instrument - agency" relations, which denote than somebody or something uses some object, and "cause - effect" relations, which represent the fact that an event or an object causes some effect. The latter finding is relevant for event coding, because news coverage often falls into this category.

I argue that the "other" class can be considered as a control case, i.e. a label for relationships that seemed relevant to human coders, but did not fit any (or maybe multiple) of the predefined categories. The results for the "other" class do not differ in any meaningful way from the results for the other classes (Table 37). This finding indicates that with respect to windowing, the specific semantic relationships considered in SemEval are representative for other types of relations and vice versa. Taking this interpretation a step further, I argue that we can generalize the insights gained about window sizes for the considered types of semantic relationship to other types of semantic relations.

Finally, I did not find any differences in window size (distribution) depending on the number of examples per relationship. This indicates that the coding guidelines used for annotation, the resulting relational data and identified effects, or both, are robust.

In ACE5, additional classes of link types are considered in comparison to SemEval; namely syntactic classes, different relationship types (similar to the semantic roles in SemEval) and subtypes, modality, and tense (Linguistic_Data_Consortium, 2005). The first two classes are relevant for this study and are discussed in detail below. Another important particularity with relations in ACE is that links can be formed between distinct entities that belong to the same extent of one entity. Such constituents are still annotated as truly distinct, individual entities in ACE. For instance, for the marked-up extent of the entity "southern Philippines airport", there is a relationship (of type "geographical') annotated between the nominals "airport" (unique entity of type "facility") and "southern Philippines" (unique entity of type "location"). For practical text coding and event coding applications, users often are often not interested in establishing links among tokens in multi-word expressions. If those relations do matter, the window size is rather deterministic, i.e. zero for adjacent terms. One goal with this project is to inform decisions about appropriate window sizes between entities that are common in texts from or about socio-technical systems. In such data, relevant mentions of entities typically do not overlap, e.g. in written accounts of who did or said what to whom in what manner. Thus, for the following analyses, it seems necessary to distinguish between relations between overlapping versus non-overlapping entities. Moreover, it seems necessary to discount for deterministic window sizes that result from overlapping entity extents as there is little new to learn about them. My analysis shows that whether the extents of linked entity mentions overlap or not is mainly a function of the syntactic class[6] of the relationship (Table 38): in ACE5, 67.5% of all links show overlaps in entity extent. Of those links, 92% are members of three syntactic classes:

- "Premod" relations, which denote links between proper adjectives or proper nouns that modify an entity, e.g. "New York police". These entities are often multi-word units that an N-gram tagger might identify as such and for which the window size would be zero.
- "Possessive" relations, where one entity is possessing the other one, e.g. "New York's citizens". These entities are often collocations, and the respective window size would also be zero.
- "Preposition" relations, where two entities are linked through a preposition, e.g. "citizens of New York". Here, the window size equals the number of tokens in the preposition, which is often one or two.

[6] In ACE, one of the intension with syntactic classes is to provide the annotators with a justification or sanity check for marking up a link.

Since the window sizes for these three types of relations are driven by syntactic rules for language production, they are not of further interest for analysis here because the respective window sizes are deterministic and can be estimated given the type of relationship. This decision does not imply that these relations are irrelevant for network analysis, in fact many of them would be true positives with many relation extraction approaches.

Table 38: Types of syntactic relationship, size in corpus, and ratio of overlapping entity extents

Syntactic Relation	Share of total dataset	Overlapping in extent
PreMod	28.2%	99.0%
Verbal	21.2%	4.9%
Preposition	19.4%	88.5%
Possessive	17.3%	98.1%
Other	8.5%	9.4%
Formulaic	3.1%	66.9%
Participial	2.0%	68.6%
Coordination	0.4%	51.6%

Table 39 provides the empiric results for the frequency and coverage rates of window sizes depending on the types of syntactic relations. "Depending on" here means given a certain window sizes; there could still be some underlying other factor that explains the observed results. The shown numbers confirm that for possessive and premod relations, the most frequent window size is zero, and over 95% of links in those classes require a window size of two or less.

In other syntactic relations, fewer entities overlap in extent: first, in "coordination" relations, where two nouns phrases are connected via the conjunction "and", e.g. "citizens and police". Most of these noun phrases are clearly distinct entities. However, the amount of words between them is still deterministic (one for "and", see Table 39 for a confirmation), and therefore are also not of interest here. Next, "formulaic relations", which mainly tie the author or reporter to a publishing location of a news article, such as in "John Doe, the BBC, London". Here, links also mainly consists of collocated entities so that the most frequent window size is zero (Table 39). Moreover, this genre-specific type of relationship cannot be assumed to generalize to other domains, and is disregarded for further analysis.

In relations of the types "participial", where a participial phrase modifies a head noun, e.g. "the people who moved to New York", and "verbal", where nodes are linked through a verb, the involved entities are typically distinct entities, and at least in the case of "verbal" also mainly non-overlapping. Moreover, links of these two types are relevant for event coding as they imply some activity (Gerner et al., 1994). With some REX approaches, verb phrases that represent activities are actually considered as nodes (Carley et al., 2007; Goldstein, 1992; King & Lowe, 2003), while in other approaches, they are not (Corman et al., 2002). Another syntactic relationship where the majority of instances do not involve overlapping entity extents is the "other" class. This is a collection of links that do not fit the definition of any of the other syntactic classes, but "beyond a reasonable doubt" are a relevant link (Linguistic_Data_Consortium, 2005). As already explained for the SemEval data, the "other" class is relevant for this study as it can serve as a control condition. Taken together, the "participial", "verbal" and "other" class account for 32.5% of all links in ACE, but only for 4.8% of the links where the extents of entities are overlapping. Based on these results and the aforementioned reasoning, I consider relations of the types "verbal", "participial", and "other" for further analysis, with the exception of the error analysis at the end of this chapter, where again all types are considered. For the considered syntactic classes (N of links =2,841), the most common window size is two or three, but it takes more than 7 (participial), 11 (verbal), or 13 (other) intervening words to identify at least 90% of the links denoted in the ground truth (Table 39).

Table 39: Impact of type of syntactic relationship on window size

Window	PreMod		Formulaic		Possessive		Coordination	
0	80.5%	80.5%	75.8%	75.8%	66.8%	66.8%	3.2%	3.2%
1	13.0%	93.5%	12.6%	88.5%	22.9%	89.6%	51.6%	54.8%
2	4.6%	98.2%	4.5%	92.9%	6.4%	96.0%	19.4%	74.2%
3	1.2%	99.4%	2.6%	95.5%	2.6%	98.6%	9.7%	83.9%
4	0.4%	99.8%	1.1%	96.7%	0.7%	99.3%	12.9%	96.8%
5	0.0%	99.9%	0.7%	97.4%	0.3%	99.6%	0.0%	96.8%
6	0.0%	99.9%	0.7%	98.1%	0.2%	99.8%	0.0%	96.8%
7	0.0%	99.9%	0.4%	98.5%	0.1%	99.9%	0.0%	96.8%
8	0.0%	99.9%	0.4%	98.9%	0.0%	99.9%	0.0%	96.8%
9	0.0%	100.0%	0.7%	99.6%	0.1%	99.9%	0.0%	96.8%
10	0.0%	100.0%	0.0%	99.6%	0.0%	99.9%	0.0%	96.8%
	Preposition		Participial		Verbal		Other	
0	1.5%	1.5%	7.6%	7.6%	3.3%	3.3%	9.4%	9.4%
1	37.3%	38.8%	11.0%	18.6%	8.6%	11.9%	8.8%	18.2%
2	31.1%	70.0%	19.8%	38.4%	15.5%	27.4%	12.9%	31.1%
3	14.9%	84.9%	20.9%	59.3%	14.7%	42.1%	10.6%	41.7%
4	6.8%	91.7%	11.6%	70.9%	13.1%	55.2%	10.5%	52.1%
5	3.5%	95.2%	8.7%	79.7%	10.3%	65.5%	8.2%	60.3%
6	1.7%	96.9%	5.8%	85.5%	7.0%	72.5%	6.6%	66.9%
7	1.0%	97.9%	5.2%	90.7%	5.5%	78.1%	6.0%	72.9%
8	0.8%	98.6%	3.5%	94.2%	4.9%	82.9%	4.6%	77.5%
9	0.5%	99.2%	1.2%	95.3%	3.2%	86.2%	4.7%	82.2%
10	0.4%	99.6%	1.2%	96.5%	3.0%	89.2%	2.7%	84.9%

The impact of genre on window size is also tested here. Table 40 lists the genres considered in this project along with their respective size in the corpus. This table also shows the ratio of the selected syntactic classes among these genres. The numbers show that syntactic relations where window sizes are fairly deterministic are more common in newswire data, while they are slightly less common in broadcast news and telephone conversations; both of which are instances of spoken language data.

Table 40: Distribution of genres across corpus and selected syntactic relations (verbal, participial, other)

Genre	All relations	Selected syntactic relations
Broadcast conversation	19.0%	18.9%
Broadcast news	23.1%	25.0%
Newswire	30.7%	23.8%
Telephone	8.5%	12.3%
Usenet	9.9%	11.3%
Weblog	8.8%	8.7%

The most common window sizes (two to three) are consistently found across all genres (Table 41). Slight exceptions are telephone conversations (about one token shorter windows than the cross-genre average), and newswire data (about one token longer). The link coverage rates depending on the window size are also very similar across genres, but only up to window size eight, where about 80% of all links are found. From there on, the window sizes needed to capture more links start to vary depending on the genre (Table 41).

Table 41: Impact of genre on window size

Window	Broadcast Conversations		Broadcast News		Newswire		Telephone		Usenet		Weblog	
0	6.3%	6.3%	5.5%	5.5%	4.1%	4.1%	4.4%	4.4%	5.4%	5.4%	5.8%	5.8%
1	8.8%	15.1%	9.8%	15.3%	7.4%	11.6%	9.7%	14.1%	9.6%	15.1%	7.4%	13.2%
2	16.7%	31.8%	13.3%	28.6%	11.4%	22.9%	25.2%	39.3%	16.7%	31.7%	10.3%	23.6%
3	14.8%	46.6%	15.0%	43.6%	13.4%	36.3%	13.5%	52.8%	10.3%	42.0%	16.5%	40.1%
4	12.3%	58.8%	13.3%	56.9%	11.4%	47.7%	11.4%	64.2%	14.7%	56.7%	10.3%	50.4%
5	10.0%	68.8%	7.4%	64.2%	9.3%	57.0%	9.7%	73.9%	10.3%	67.0%	15.3%	65.7%
6	6.3%	75.1%	6.8%	71.0%	8.4%	65.3%	5.3%	79.2%	7.1%	74.0%	5.8%	71.5%
7	5.4%	80.5%	6.3%	77.3%	5.3%	70.7%	6.7%	85.9%	3.8%	77.9%	5.8%	77.3%
8	4.6%	85.1%	5.1%	82.4%	5.5%	76.1%	3.2%	89.1%	4.2%	82.1%	4.5%	81.8%
9	3.3%	88.3%	3.6%	86.0%	4.3%	80.4%	2.9%	92.1%	3.5%	85.6%	2.5%	84.3%
10	2.5%	90.8%	3.8%	89.8%	3.3%	83.7%	1.5%	93.5%	2.9%	88.5%	1.2%	85.5%
11	2.3%	93.1%	1.6%	91.3%	1.8%	85.6%	1.5%	95.0%	1.6%	90.1%	2.5%	88.0%
12	0.8%	93.9%	2.2%	93.5%	2.6%	88.1%	1.8%	96.8%	3.2%	93.3%	2.5%	90.5%
13	1.5%	95.4%	2.0%	95.5%	1.7%	89.8%	0.9%	97.7%	1.3%	94.6%	2.5%	93.0%
14	0.6%	96.0%	1.4%	97.0%	1.1%	90.9%	1.2%	98.8%	1.3%	95.8%	0.8%	93.8%
15	1.0%	96.9%	0.7%	97.7%	2.3%	93.2%	0.3%	99.1%	0.3%	96.2%	0.8%	94.6%
16	1.3%	98.3%	0.3%	98.0%	0.8%	93.9%	0.3%	99.4%	0.3%	96.5%	1.2%	95.9%

17	0.4%	98.7%	0.3%	98.3%	0.8%	94.7%	0.3%	99.7%	0.6%	97.1%	0.0%	95.9%
18	0.0%	98.7%	0.1%	98.4%	0.5%	95.1%	0.0%	99.7%	0.6%	97.8%	0.0%	95.9%

In addition to testing the window size depending on genre, window sizes are also analyzed herein depending on the following types of relationships, which are conceptually similar to the semantic relations in SemEval:

- Social, personal: relations between people.
- Organizational affiliation: professional relations, such as employment.
- General affiliation: relations between people and organizations in the widest sense or geopolitical entities, e.g. residency or religion.
- Agent-Artifact: social agent owning an artifact.
- Physical: the location of a person.
- Part whole: the location of objects, and hierarchical relations among and between social agents and objects.

The share of each of these types of relationships in the dataset and among the selected syntactic relations is shown in Table 42. Grammatically induced window sizes are prevalent in all but the geo-physical and to a lesser degree also in the agent-artifact relations. The results about window size per semantic relationship based on SemEval (Table 43) confirm the previous findings based on ACE: typical window sizes (two or three) and coverage rates are very similar across the different types of relationships. The "part-whole" relationship requires a slightly shorter distance, and the same had been observed for the "component-whole" type in ACE. However, when filtering the links in ACE depending on the type of semantic relationship as done in this study, the average link coverage rates in ACE lag behind the rates found in SemEval. One explanation for this difference might be that in ACE, I eliminated certain grammatical relationships because their window size is deterministic (driven by rules for language production) and already know. This was not possible for SemEval since no syntactic classification of links is provided in this corpus. However, a closer inspection of the links with low window size in SemEval suggested that these also represent grammatical dependencies. Therefore, the links in SemEval are a mixture of short, mainly grammatically motivated relations and other types of relations that are of stronger interest here. In ACE, I was able to separate these types of relationships more precisely; showing that the type of grammatical relationship (or lack thereof, as in the "other" type), has a major impact on window size.

Table 42: Types of semantic relationships, size in corpus, size among selected syntactic relations

Type	All relations	Selected syntactic relations
Agent-Artifact	10.0%	14.2%
General affiliation	11.0%	5.5%
Organizational affiliation	29.0%	13.8%
Part Whole	14.9%	4.3%
Personal and social	12.5%	7.9%
Physical	22.6%	54.3%

Table 43: Impact of type of semantic relationships on window size

Window	Personal, social		Organizational affiliation		General affiliation		Agent Artifact		Part Whole		Physical		Average	
0	3.6%	3.6%	7.1%	7.1%	10.5%	10.5%	1.8%	1.8%	7.6%	7.6%	5.1%	5.1%	6.0%	6.0%
1	9.5%	13.2%	7.3%	14.4%	9.2%	19.7%	7.9%	9.7%	5.9%	13.6%	9.5%	14.6%	8.2%	14.2%
2	17.3%	30.5%	11.5%	26.0%	36.2%	55.9%	17.3%	27.0%	11.0%	24.6%	13.2%	27.9%	17.8%	32.0%
3	10.9%	41.4%	17.3%	43.3%	9.9%	65.8%	15.8%	42.7%	12.7%	37.3%	13.6%	41.5%	13.4%	45.3%
4	11.4%	52.7%	12.6%	55.9%	9.9%	75.7%	14.8%	57.5%	12.7%	50.0%	12.0%	53.5%	12.2%	57.5%
5	10.9%	63.6%	12.6%	68.5%	5.3%	80.9%	8.9%	66.4%	9.3%	59.3%	9.3%	62.8%	9.4%	66.9%
6	4.5%	68.2%	4.5%	73.0%	3.3%	84.2%	7.6%	74.0%	7.6%	66.9%	7.8%	70.6%	5.9%	72.8%
7	8.2%	76.4%	5.2%	78.2%	3.3%	87.5%	5.9%	79.9%	5.9%	72.9%	5.5%	76.1%	5.7%	78.5%
8	5.9%	82.3%	4.7%	82.9%	2.6%	90.1%	5.1%	85.0%	4.2%	77.1%	4.7%	80.8%	4.5%	83.0%
9	4.5%	86.8%	2.4%	85.3%	2.6%	92.8%	2.3%	87.3%	6.8%	83.9%	3.8%	84.6%	3.7%	86.8%
10	1.8%	88.6%	3.7%	89.0%	1.3%	94.1%	2.0%	89.3%	5.1%	89.0%	2.9%	87.5%	2.8%	89.6%
11	1.4%	90.0%	2.1%	91.1%	0.0%	89.0%	1.5%	90.8%	0.0%	89.0%	2.3%	89.8%	1.2%	89.9%
12	1.4%	91.4%	2.4%	93.4%	2.0%	96.1%	1.8%	92.6%	3.4%	92.4%	2.1%	91.9%	2.2%	93.0%
13	0.9%	92.3%	1.3%	94.8%	1.3%	97.4%	2.0%	94.7%	0.0%	92.4%	1.9%	93.8%	1.3%	94.2%
14	2.7%	95.0%	0.3%	95.0%	0.7%	98.0%	0.8%	95.4%	2.5%	94.9%	1.1%	94.9%	1.3%	95.5%
15	1.8%	96.8%	0.5%	95.5%	0.0%	98.0%	1.0%	96.4%	1.7%	96.6%	1.1%	96.0%	1.0%	96.6%

Most of the types of semantic relationships in SemEval and partially also in ACE are defined over entity types, i.e. they can only be established between certain node classes. In this sense, semantic relationships are a proxy for the impact of the node class or classes involved in a link on window size. This impact can be determined even more precisely by analyzing the window size for all combinations of node classes considered in SmeEval[7]. Table 44 shows how these types of links are distributed across the corpus; indicating that the vast majority of links (over 85%) occur between a person and a) another person (7.5% of all links) or b) some other entity class (77% of all links). Only four percent of all links do not involve a social agent (person or organization). Therefore, the findings from this analysis are highly relevant for constructing social network data that involve people and organizations, and socio-technical network data (social agents linked to some other entity type). Looking at window size in dependence of node classes involved in links, again, the common window sizes and coverage rates are highly similar across node classes and to the previous findings (Table 45). The exceptions are "person-time" relations, where the window size is about two tokens longer than for the other types, and "location-location" relations, which are shorter than the average by about one token. Looking at aggregated groups of node classes with respect to link coverage rates, the results suggest that the rates grows fastest for spatial relations (window sizes here are comparatively shorter than for the other groups, size 10 for 90% of the links); followed by relations between social agents and resources (Table 45). For relations between social agents only, average window sizes are comparatively longest (12 for 90% of the links). However, these differences are still small.

Table 44: Links per entity class

Entity Class	Person	Organization	Location	Resource	Time
Person	7.5%	18.7%	34.9%	6.6%	16.8%
Organization	0.5%	2.5%	1.7%	3.6%	0.7%
Location	1.4%	0.7%	3.6%	0.0%	0.0%
Resource	0.3%	0.0%	0.0%	0.3%	0.0%
Time	0.0%	0.0%	0.0%	0.0%	0.0%

[7] The entity classes in ACE are: person, organization, geopolitical entity (GPE), location, facility, vehicle, and weapon. In order to keep the findings comparable to further analyses on the node class level (chapters 4 and 5), I mapped the ACE classes to the meta-network classes as follows: Agent: person. Organization: organization and GPE except for population center and state. Location: location, GPE (except for country, GPE cluster, nation, continent, special) , and facility. Resource: vehicle and weapon.

Table 45: Impact of entity class on window size*

Win	Person		Person		Person		Person		Person	
dow	Person		Organization		Location		Resource		Time	
0	3.4%	3.4%	5.6%	5.6%	5.1%	5.1%	2.1%	2.1%	7.1%	7.1%
1	8.7%	12.0%	9.8%	15.4%	7.9%	12.9%	7.8%	9.9%	11.0%	18.1%
2	18.8%	30.8%	15.4%	30.8%	16.7%	29.7%	16.1%	26.0%	7.1%	25.2%
3	11.5%	42.3%	16.0%	46.8%	15.1%	44.8%	19.8%	45.8%	8.4%	33.5%
4	11.5%	53.8%	11.1%	57.9%	13.6%	58.5%	14.1%	59.9%	9.9%	43.4%
5	11.1%	64.9%	11.7%	69.5%	7.9%	66.3%	8.9%	68.8%	10.3%	53.8%
6	4.3%	69.2%	6.0%	75.6%	8.1%	74.4%	7.3%	76.0%	6.2%	60.0%
7	7.2%	76.4%	5.1%	80.6%	5.9%	80.3%	5.7%	81.8%	6.0%	66.0%
8	5.8%	82.2%	4.7%	85.3%	3.7%	84.0%	4.7%	86.5%	6.5%	72.5%
9	4.3%	86.5%	2.4%	87.8%	3.8%	87.7%	0.5%	87.0%	3.9%	76.3%
10	1.9%	88.5%	3.2%	91.0%	2.2%	89.9%	2.1%	89.1%	4.3%	80.6%
11	1.4%	89.9%	1.3%	92.3%	2.3%	92.2%	1.0%	90.1%	3.2%	83.9%
12	1.4%	91.3%	2.6%	94.9%	1.9%	94.1%	2.6%	92.7%	2.2%	86.0%
13	1.0%	92.3%	1.5%	96.4%	1.7%	95.8%	1.6%	94.3%	2.6%	88.6%
14	2.9%	95.2%	0.8%	97.2%	0.7%	96.5%	1.0%	95.3%	1.1%	89.7%
15	1.4%	96.6%	0.4%	97.6%	0.9%	97.4%	0.5%	95.8%	2.4%	92.0%
	Organization		Organization		Organization		Location		Average	
	Organization		Resource		Location		Location		(unweighted)	
0	2.9%	2.9%	1.0%	1.0%	9.1%	9.1%	5.9%	5.9%	4.7%	4.7%
1	4.3%	7.2%	6.0%	7.0%	10.6%	19.7%	7.9%	13.9%	8.2%	12.9%
2	20.3%	27.5%	24.0%	31.0%	15.2%	34.8%	12.9%	26.7%	16.3%	29.2%
3	13.0%	40.6%	10.0%	41.0%	13.6%	48.5%	16.8%	43.6%	13.8%	43.0%
4	10.1%	50.7%	11.0%	52.0%	15.2%	63.6%	11.9%	55.4%	12.0%	55.0%
5	10.1%	60.9%	9.0%	61.0%	10.6%	74.2%	11.9%	67.3%	10.2%	65.2%
6	4.3%	65.2%	9.0%	70.0%	4.5%	78.8%	7.9%	75.2%	6.4%	71.6%
7	5.8%	71.0%	6.0%	76.0%	0.0%	78.8%	5.9%	81.2%	5.3%	76.9%
8	4.3%	75.4%	4.0%	80.0%	4.5%	83.3%	5.0%	86.1%	4.8%	81.7%
9	5.8%	81.2%	7.0%	87.0%	6.1%	89.4%	3.0%	89.1%	4.1%	85.8%
10	5.8%	87.0%	3.0%	90.0%	1.5%	90.9%	3.0%	92.1%	3.0%	88.8%
11	0.0%	87.0%	1.0%	91.0%	0.0%	90.9%	0.0%	92.1%	1.1%	89.9%
12	1.4%	88.4%	1.0%	92.0%	3.0%	93.9%	3.0%	95.0%	2.1%	92.1%
13	0.0%	88.4%	4.0%	96.0%	0.0%	93.9%	0.0%	95.0%	1.4%	93.4%
14	5.8%	94.2%	0.0%	96.0%	0.0%	93.9%	1.0%	96.0%	1.5%	94.9%
15	1.4%	95.7%	1.0%	97.0%	0.0%	93.9%	1.0%	97.0%	1.0%	95.9%

* Only type of entity to entity connections with 20 or more links considered. Relations are directional in the data. Here, both directions are taken together per type.

Table 46 provides a brief summary of the results from the window size analysis reported in this chapter. This synopsis shows that after controlling for the type of syntactic relationship, i.e. excluding relationships where the window sizes are short and deterministic due to syntactic rules of language production, there are virtually no differences between typical window sizes and link coverage rates across genres, types of syntactic relationships, types of semantic relationships, and types of node classes involved in links.

Table 46: Summary of results for windowing

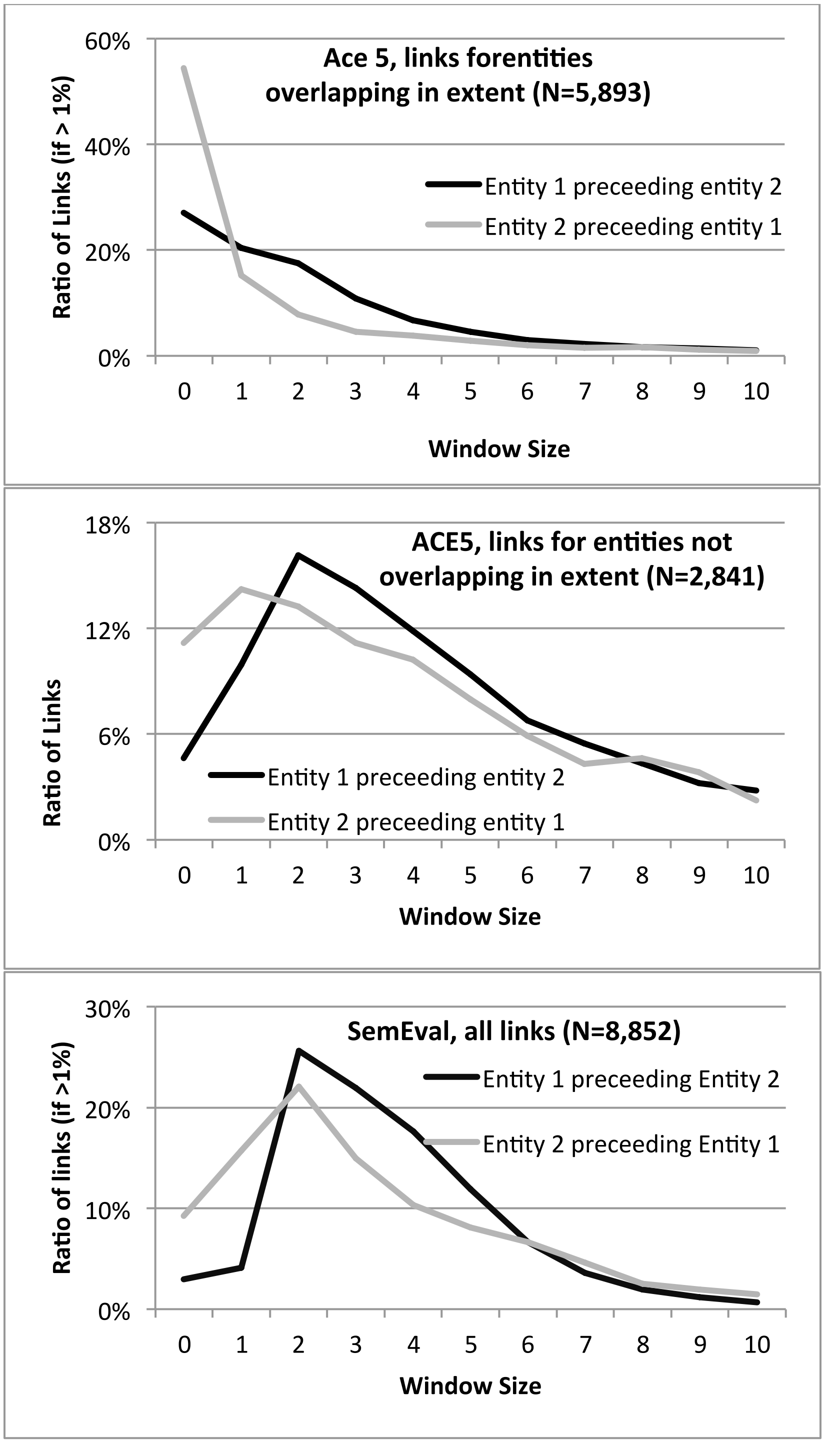
		SemEval	ACE5			
		Semantic relations	Syntactic relations	Semantic* relations	Node class*	Genre*
Most frequent window size		2	2	2	2	2 and 3
Link coverage rate	50%	3	4	4	4	4
	75%	5	7	7	7	7
	80%	5	8	8	8	8
	90%	7	10	12	12	11
	95%	8	13	14	14	14

* controlled for type of syntactic relation (only including verbal, participial, other)

Finally, the data show an impact of entity ordering on window size: in more than half of all links, the first entity in a relationship precedes the second one (55% of all links in SemEval[8], 58% in ACE). If this is the case, the average window size is about one word longer than when the second entity precedes the first one (Figure 8). This ordering effect disappears at about window size six, and is also similar across all types of relationships and nodes in links for both corpora. The results in Figure 8 also show that for linked entities with non-overlapping extents (ACE), the patterns of link coverage rates depending on window size are highly similar for both corpora. This holds true even though these two corpora differ considerably in genres, time of data collection, and types of entities and relations considered. Therefore, this finding suggests that the presented results for typical window sizes and the amount of links identified depending on window size are highly robust across genres, time, data sources, and types of relationships. This implies that the window sizes found with this study are likely to generalize to other text data.

[8] The analysis of order effects excludes the “other” relationship because no entity order is marked up for these relations.

Figure 8: Impact of ordering effects on window size and link coverage



2.7.2.2 Evaluation of windowing

Using windowing for connecting nodes into edges implies the danger of missing links (false negatives) and retrieving incorrect links (false positives). This potential cause of errors has been repeatedly pointed out in the past (Carley, 1997a; Corman et al., 2002), but has not yet been empirically tested. I am quantifying the amount of these errors based on the SemEval and ACE5 data.

The results show that the rates of false negatives decline rapidly; falling below 5% at window size 8 (SemEval) to 9 (ACE, text-level, all types of relations considered). At window size 12, the rate of false negatives is less than 2.4% (ACE5) to 1% (SemEval) (Table 47, Table 48, Table 49). Table 47 and Table 48 express these errors in terms of false positives and false negatives, and Table 49 represents the same errors in terms of recall, precision, and the harmonic mean of these two metrics (F).

Table 47: Accuracy rates and false negatives due to windowing (SemEval)

Window Size	Correct	False Negatives
0	5.9%	94.1%
1	15.2%	84.8%
2	38.8%	61.2%
3	57.8%	42.2%
4	72.3%	27.7%
5	82.5%	17.5%
6	89.1%	10.9%
7	93.2%	6.8%
8	95.4%	4.6%
9	97.0%	3.0%
10	97.9%	2.1%
11	98.6%	1.4%
12	99.1%	0.9%

The rate of false positives was measured by connecting the heads of any nodes that are annotated as entities in the ground truth data if the number of tokens between these heads is equal to or lower than a given window size. This was done for ACE, but could not be done for SemEval because there, only two entities are marked up per sentence, and the sentences are not consecutive. Links in ACE are mainly marked up within sentences. However, 4.2% of all links span across sentences. For real world applications, considering cross-sentence links can be an appropriate approach, e.g. when an event is described over multiple sentences[9]. In order to

[9] In order to accommodate for that in AutoMap, users there can chose the number of sentences after which the window should be reset.

clarify on the impact of distinguishing between within versus across sentence links, I show the results for both scenarios in Table 48 and Table 49: for the lower halves of these tables, windows were reset at the end of sentences. A side effect of this distinction is that with the sentence level approach, the rate of false negatives (7.2% at window size 12) will be higher since some links cannot be found within sentences. Sentences splitting was conducted by considering each dot as a sentence mark unless the dot occurs right next to a list of 86 terms (e.g. Dr., D.C.) that I identified by manually checking all actual cross-sentence links in ACE. This way of sentence splitting is on the conservative side, i.e. there might be more sentences identified than there really are. I chose this approach to make sure that the number of false positives is not overestimated. Therefore, my results show the lower bound of false positives due to windowing in addition to the more unconstrained, cross-sentence setting.

Overall, the rate of false positives is alarmingly high. When considering all additional links retrieved, the rate of false positives is similar to the rate of correctly identified links. For example, at window size 7, 88.9% (sentence level) to 92.5% (cross-sentence level) of false positives are returned (Table 48, 4th column). This means that when a window size of 7 is applied, 9 out of 10 of the retrieved links were not annotated by human coders as being relevant.

Further analyzing the false positives revealed that in many cases, the many of the involved entities were overlapping. As mentioned previously in this chapter, such entities often represent regular multi-word expressions, e.g. “UN Security Council”, or consist of a named entity plus a role or attribute of the entity, e.g. “Palestinian security sources”. However, for REX purposes, users would typically not create links within meaningful N-grams, and roles and attributes are often not considered as a node class of their own, but only as attributes of nodes. Therefore, I conducted a second analysis of false positives were I excluded any links between overlapping entity extents from counting false positives. This experimental condition is referred to as restriction 1 in Table 48 and Table 49. After applying this restriction, the remaining false positives contained a large number of entities from the node class “time” (timex), such as dates and clock time. Since these entities never have a head but only an extent, which could span more tokens that the heads of other entities, I also excluded the timex entities in restriction 1. Another sizable portion of entities involved in false positives were references to media organizations, which typically occur at the beginning or end of news articles. Since these entities are atypical in genres other than news data, they were also disregarded in restriction 1. Overall, applying restriction 1 lowers the number of false positives per window size by thousands of links. However, at window size 7, there are still 84.6% to 90.0% of links that are false positives (Table 48).

Further analyzing the remaining false positives showed that many entities involved were pronouns. Therefore, I introduced restriction 2, which assumes that anaphora resolution had been applied prior to relation extraction as follows: pronouns get translated into entities that are referred to by a name or nominal, and a legitimate link from such an entity to another entity already exists, such that the false positive would only increase the weight of an existing link. For details on the impact of anaphora resolution on network data see the previous results section. This is a very optimistic assumption, and it is meant to show the lower bound for false positives due to windowing, even though this might be an underestimation. Applying restriction 2 in addition to restriction 1 further cuts down the rate of false positives to less than the rate of correct links, but the false positives still exceeds 68.9% to 84.6% at window size 7, and further increase from there on (Table 48). Further inspecting the remaining false positives suggested that these were not connections between named entities and roles or attributes associated with these entities. Also, the remaining false positives did not seem to be instances of any other types of meaningful relations that were emerging or discovered from the data, but rather random connection between nearby entities that did not seem obviously reasonable.

The results in Table 49 show that when using windowing, recall is acceptably high - over 90% from window size 6 (cross-sentence level) to 9 (sentence level) on. Note that recall is not impacted by applying the restrictions explained in the previous paragraph. However, the harmonic mean of recall and precision is fairly low due to the low precision rates; not exceeding 18% at window size 7.

Table 48: Error rates for windowing I (ACE)

Window Size	Correct	False Negatives	False Positives		
			All	Restriction 1	Restriction 2
	Text level (resembling ground truth)				
0	38.6%	61.4%	55.3%	36.6%	19.2%
1	56.7%	43.3%	73.4%	60.7%	37.2%
2	70.2%	29.8%	81.1%	73.1%	52.0%
3	78.3%	21.7%	85.4%	79.6%	61.1%
4	83.9%	16.1%	88.1%	83.7%	58.2%
5	87.7%	12.3%	90.0%	86.5%	72.3%
6	90.3%	9.7%	91.4%	88.5%	76.0%
7	92.4%	7.6%	92.5%	90.0%	78.8%
8	94.0%	6.0%	93.3%	91.1%	81.0%
9	95.2%	4.8%	94.0%	92.1%	82.8%
10	96.3%	3.7%	94.5%	92.8%	84.3%
11	96.9%	3.1%	95.0%	93.4%	85.5%
12	97.6%	2.4%	95.3%	94.0%	86.6%

	Sentence level				
0	35.3%	64.7%	48.0%	26.5%	11.9%
1	53.0%	47.0%	67.6%	52.8%	28.1%
2	66.1%	33.9%	76.2%	65.1%	40.5%
3	74.1%	25.9%	81.0%	72.6%	50.0%
4	79.6%	20.4%	84.0%	77.3%	56.6%
5	83.3%	16.7%	86.2%	80.5%	61.8%
6	85.8%	14.2%	87.7%	82.9%	65.8%
7	87.8%	12.2%	88.9%	84.6%	68.9%
8	89.4%	10.6%	89.8%	85.9%	71.2%
9	90.5%	9.5%	90.5%	87.0%	73.1%
10	91.5%	8.5%	91.1%	87.8%	74.6%
11	92.1%	7.9%	91.5%	88.5%	76.0%
12	92.8%	7.2%	91.9%	89.1%	77.1%

Table 49: Error rates for windowing II (ACE)

Window Size	Recall	All false positives		Restriction 1		Restriction 2	
		Precision	F	Precision	F	Precision	F
	Text level (resembling ground truth)						
0	38.6%	17.3%	23.8%	24.4%	29.9%	31.2%	34.5%
1	56.7%	15.1%	23.8%	22.3%	32.0%	35.6%	43.7%
2	70.2%	13.3%	22.3%	18.9%	29.8%	33.7%	45.6%
3	78.3%	11.4%	20.0%	16.0%	26.6%	30.5%	43.9%
4	83.9%	10.0%	17.8%	13.7%	23.5%	35.1%	49.5%
5	87.7%	8.7%	15.9%	11.9%	20.9%	24.3%	38.1%
6	90.3%	7.7%	14.2%	10.4%	18.6%	21.7%	34.9%
7	92.4%	6.9%	12.9%	9.2%	16.8%	19.6%	32.3%
8	94.0%	6.3%	11.8%	8.3%	15.3%	17.9%	30.1%
9	95.2%	5.7%	10.8%	7.6%	14.0%	16.4%	28.0%
10	96.3%	5.3%	10.0%	6.9%	12.9%	15.1%	26.2%
11	96.9%	4.9%	9.3%	6.4%	11.9%	14.0%	24.5%
12	97.6%	4.5%	8.7%	5.9%	11.1%	13.1%	23.0%
	Sentence level						
0	35.3%	18.3%	26.0%	31.1%	24.1%	29.9%	33.1%
1	53.0%	17.2%	25.0%	38.1%	25.9%	34.0%	44.3%
2	66.1%	15.7%	23.1%	39.3%	25.4%	34.2%	49.3%
3	74.1%	14.1%	20.3%	37.1%	23.7%	31.9%	49.4%
4	79.6%	12.7%	18.1%	34.5%	21.9%	29.5%	48.1%
5	83.3%	11.5%	16.2%	31.8%	20.3%	27.2%	46.1%
6	85.8%	10.5%	14.7%	29.3%	18.8%	25.1%	43.7%
7	87.8%	9.8%	13.5%	27.3%	17.6%	23.4%	41.7%
8	89.4%	9.1%	12.6%	25.7%	16.6%	22.0%	40.0%
9	90.5%	8.6%	11.8%	24.4%	15.7%	20.9%	38.4%
10	91.5%	8.2%	11.1%	23.2%	15.0%	19.9%	37.0%
11	92.1%	7.8%	10.6%	22.1%	14.4%	19.0%	35.7%
12	92.8%	7.5%	10.1%	21.3%	13.8%	18.3%	34.6%

2.7.2.3 Windowing: Answers to research questions

The empirical results from the windowing study suggest the following answers to the research questions:

5. Question: *What window size do human experts use when identifying relations in text data?*

Does this typical window size differ depending on the type of data or relations?

5. Answer: Regardless of text genre and the type of semantic relationship, syntactic relationship, and node classes, the most frequently used window size is two.

6. Question: *What window size is needed to capture the vast majority of links in text data? Does this window size differ depending on the type of data or relations?*

6. Answer: On average and regardless of text genre and the type of semantic relationship, syntactic relationship, and the classes of nodes involved in a link, at least 50% of all links are found when using a window size of four. After that, window sizes vary depending on the type of syntactic relationship: for mainly syntactically motivated relations, it is sufficient to choose a window size of four to retrieve over 90% of the links. Excluding these syntactic relations, a window of at least twelve is needed to achieve the same result. If a corpus contains an indistinguishable mixture of both types of links; at least 90% of all links are covered with a window size of seven. After controlling for the type of syntactic relationships, i.e. excluding relationships where the window size is short and deterministic due to syntactic rules of language production, these findings are robust across text genres, types of semantic relationships, and node classes. In summary, meaningful differences between link coverage rates are due to syntactic relations. Finally, window sizes also differ depending on ordering effects of the occurrence of entities in the text data. The latter effect is also robust across the test corpora.

7. Question: *What error rate, i.e. amount of wrongfully identified links (false positives) and missed links (false negatives), can be expected when applying a specific window size? Does this error rate differ depending on the type of data or relations?*

7. Answer: Based on the ground truth datasets used herein, the rate of false negatives declines rapidly; falling below 5% at window size eight to nine. At window size twelve, the rate of false negatives is 2.4% (excluding certain abovementioned syntactic relations) to less than 1% (incl. those syntactic relations). However, the rate of false positives is alarmingly high: when coding links across sentences, the rate of false positives ranges between 79% to 93% at window size seven, and 87% to 95% at window size twelve. When coding links only within sentences, the rate of false positives varies between 69% to 89% at window size seven, and 77% to 92% at window size 12. The variances in range are due to eliminations of certain types of entities involved in false positives. Therefore, the presented results can be interpreted as an empirically grounded upper bound and lower bound for the rates of false positives due to windowing.

2.8 Conclusions

The results from the reference resolution project and the windowing project show that the coding choices that need to be made when extracting entities and relational data from texts strongly impact the network properties and structure. The conclusions from the experimental work are presented in this section. The practical implications of the findings from this chapter for applied work are synthesized in chapter 4.

The goal with RR is to map pronouns and additional entity mentions to the set of unique entities; thereby reducing the amount of pronouns and unassociated entities while increasing the weight per unique entities. The results from the RR study indicate that the deduplication, consolidation and personalization of entities has a strong impact on the node, link and network level, especially with respect to quantitative analysis results: applying both, AR and CR, alters the identity and weight of about 76% of all entity mentions, and the average weight per unique entity or node is increased from 1.0 to 5.8. As a result, less than 18% of the unique nodes carry more 79% of the total node weight. The impacts are less strong on the link level: In about 23% of all links, at least one node is changed due to AR, and 6% of all links are reduced via CR. Combining both techniques leads to a link reduction of 12%. Of the remaining links, 11% are changed due to RR, and they carry 23% of the total link weight. On the network level, the values of several metrics change strongly when applying RR, for example degree centralization, clustering coefficients, and connectedness (all increased), while a smaller number of metrics is not impacted, e.g. fragmentation, efficiency and hierarchy. In comparison to the raw data, the set of key players identified through network analysis completely changes when applying AR and CR; with CR having a stronger impact on the outcome. For all observed effects, combining AR and CR is more effective than applying either technique alone.

The ratios of resolvable anaphora as well as entities that can be co-referenced are similar across all genres considered. However, the impact of either technique on a corpus from a given domain varies depending on the distributions of pronouns, names, and nominal: in newswire and newspaper data, names and nominals are dominating, and therefore, CR is more effective than AR. In telephone conversations, where pronouns are dominating, AR makes a bigger difference than CR does. In social media data, the difference in the effectiveness per technique is more balanced, and both techniques together are highly effective (74% of entities changed).

The findings from simulating the impact of typical error rates for RR on changes in the resulting network data show that the amount of change in the value of network analytical metrics by far exceeds the change rate in RR accuracy (for 13 of 20 measures tested). The set of the nodes that score highest on these metrics is more robust towards changes in RR accuracy.

The results from the impact of windowing on link formation show that expert human coders typically apply short window size, which are mainly two to three words long. A window size of twelve is sufficient to identify more than 90% of all links in the ground truth data. These findings are robust: after disregarding relationships where the window sizes are deterministic due to syntactic rules of language production, there are virtually no differences between typical window sizes and link coverage rates across different datasets, genres, types of syntactic relationships, types of semantic relationships, and types of node classes involved in links.

The error analysis of links found by using windowing revealed that the amount of false negatives (missing links) is low; falling below 5% at window sizes eight to nine. However, the rate of false positives (additional links retrieved) is alarmingly high; reaching 90% at window size five. The rate of false positives shrinks when corpus-specific peculiarity of annotating entities and relations are disregarded, but still reaches 90% at window size seven. Assuming that AR would have been applied to the data such that no pronouns are left in any link further reduces the rate of false positives to 87% at window size twelve.

2.9 Limitations and Future Work

The insights gained with the reference resolution study and the windowing study strongly depend on the data. Even though multiple datasets were reviewed for their eligibility for this study, and multiple datasets have been analyzed, other data might have lead to different results, or provide further support for the presented findings.

The findings on the joint impact of AR and CR are furthermore limited by the order of the application of these routines. I used AR prior to CR, and this reflects common practice. With this approach, the amount of non-pronominal entities is increases first, which can then be exploited by CR. However, performing CR first might result in a less confusing mass of entities to choose from for AR. Further work is needed to identify the optimal ordering of AR and CR.

One could argue that the shown differences in the values of network analytical measures depending on RR techniques are influenced by the size of the network. In fact, prior research has shown how robust certain network metrics towards missing data and thus network size (Borgatti, Carley, & Krackhardt, 2006; Frantz, Cataldo, & Carley, 2009). However, the RR techniques impact the network size in the first place. Therefore any identified changes might still correlate with changes in network size, but the driving underlying mechanism is still the applied RR techniques.

The RR study has shown how RR techniques help to bring network data extracted from texts closer to the true underlying network structure. A valuable extension to this work, which could also help to improve reference resolution from an NLP point of view, would be to use network analysis to identify the structural position and properties of nodes on which reference resolution would be most effective. This might be useful, for example, for frequently mentioned pronouns as well as nodes representing agents with common names that need to be split up into truly distinct individuals. Here, AR and CR could be applied to separate highly central yet almost generic nodes, such as “they” or “Smith”, into multiple and distinct pronouns and nominal, respectively. The question here is: are the properties of these nodes distinct from other nodes and can thus be identified with network analysis? The outcome of such an extension could be a mechanism that suggests nodes for further treatment with RR to the user.

Finally, two preprocessing techniques and one link formation technique that are applicable when coding texts as networks were investigated. These techniques were selected because they are commonly used. Moreover, co-reference resolution and windowing are available in AutoMap, but we did not have a clear understanding of their impact on the networks extracted with AutoMap. In order to gain a more comprehensive understanding of the impact of coding choices on network data and analysis results, more techniques need to be investigated, especially alternative link formation approaches, such as techniques based on syntax and semantics of text data.

3 Computational Integration of Network-Centric Classification Model and Supervised Machine Learning for Entity Extraction

One key step in Relation Extraction (REX) is the extraction of entities from text data, which are then used as nodes for constructing network data (McCallum, 2005). Extracting entities from texts also exists as a standalone task, which is referred to as Entity Extraction (Bikel, Miller, Schwartz, & Weischedel, 1997). This chapter makes a contribution by developing an answer and computational solution to the following question:

- How can we build an entity extractor as part of a relation extraction system that supports end-users in analyzing networks and addressing substantive questions about socio-technical networks?

This chapter is structured as follows: first, the set of requirements for an entity extractor is identified (3.2.2). The various methods that are available for conducting entity extraction are then reviewed (3.2.3), and the most suitable one with respect to the identified requirements is selected. Next, I describe how I adapted and further advanced a technology that implements the selected method (3.3) and report on the performance of the resulting technology (3.4).

3.1 Introduction and Problem Statement

Entities or nodes extracted from text data are referred to as concepts, which are abstract representations of what people conceive in their minds (Sowa, 1984). Methods for entity extraction differ depending on what type of network data need is needed: For generating one-mode networks from texts, it is sufficient to correctly locate the relevant entities in the text data, a task also referred to as boundary detection, and then linking them into edges (Carley, 1994; Danowski, 1993). The resulting networks are often called concept networks, and sometimes also semantic networks (Diesner & Carley, 2011a). To keep terminology coherent in this document, I refer to relational representations of language and knowledge as "concept networks" (for a brief synopsis see Diesner & Carley, 2010c; Diesner & Carley, 2011b). One-mode concept networks have been typically used to answer questions like: What concepts, topics or memes emerge, spread and vanish in socio-technical networks? How do such diffusion processes happen? (Corman et al., 2002; Doerfel & Barnett, 1999; Gloor et al., 2009; Griffiths et al., 2007; Leskovec et al., 2009) Sometimes, the nodes in such networks are further connected to nodes representing the agents who have generated the information represented by the concept nodes or to the documents in which this information occurs. Such networks are often constructed as bipartite graphs, and have been used to address questions like: Who is talking to whom about what? Who is setting what trends? Who is an expert on which topic? (Ehrlich, Lin, & Griffiths-

Fisher, 2007; Giuffre, 2001; Gloor & Zhao, 2006; C. Roth & Cointet, 2010; Shahaf & Guestrin, 2010) For building multi-mode networks, the located entities need to be further assigned to entity classes, which are also known as categories. This assignment typically happens according to some ontology, which can be predefined or derived from the data (Van Atteveldt, 2008). State of the entity extraction and relation extraction technologies typically facilitate the retrieval of named and unnamed mentions of the following entity classes: people, organizations, locations and miscellaneous or other entities (Borthwick, Sterling, Agichtein, & Grishman, 1998; Schrodt, 2001). The resulting network data have been used to address questions like: Who is talking to whom? Who are the key players in a group? What opportunities and challenges result from the observed structure and properties of a network for an organization or a social system? (Carley et al., 2007; Hämmerli et al., 2006; Van Atteveldt, 2008)

Accuracy rates for NER systems have steadily increased over the last decade; currently being in the 80ies and lower 90ies for English (see for example Florian, Ittycheriah, Jing, & Zhang, 2003). Since such systems often focus on the extraction of entities that are referred to by a name, this process is also called Named Entity Recognition (NER) (Bikel et al., 1997; Klein, Smarr, Nguyen, & Manning, 2003; Ratinov & Roth, 2009). In NLP and political science, the default set of types of named entities to extract has remained fairly unchanged over the last decade. However, for studying the properties and functioning of socio-technical networks and addressing substantive questions about networks and their context, the classic set of named entity classes might not suffice: in addition to knowing which social agents and locations are relevant and connected, one might also need relational data about the what (*tasks* and *events*), how (*resources* and *knowledge*), why (*beliefs* and *sentiments*) and when (*time*) of interactions and activities (Barthelemy et al., 2005; Carley, 2002a). Since mentions of instances of these additional entity classes are often not referred to by a name, I refer to the more general task of extracting named *and* unnamed entities as "entity extraction". Entity Extraction allows for the construction of richer multi-mode data than NER does. The data resulting from Entity Extraction allow for moving beyond asking questions about social networks, other types of one-mode networks, and bipartite graphs in which one type of nodes are social agents, to also address questions like: Which tasks and events are some social agents involved in? What resources and knowledge are at some social agents' disposal? What impact does resource allocation have on task completion? What is the interplay of social and technical structures, and how do these structures co-evolve? (Carley, 2002a; Cataldo, Wagstrom, Herbsleb, & Carley, 2006; Krackhardt & Carley, 1998) Finally, for sentiment analysis and social media analysis - two subareas of Information Extraction that are currently highly popular and gaining further momentum - such additional categories are essential for analyzing individual and collective behavior (see for example Qureshi, Memon, Wiil, & Karampelas; Whitelaw, Patrick, & Herke-Couchman, 2006).

Looking at NER tools from the perspective of end-users who want to analyze socio-technical network data, there is another shortcoming: from an NLP perspective, efforts in advancing NER have been focused on improving the accuracy and efficiency of extractors, while transitioning from learned and evaluated models to readily usable end-user technologies has gotten less attention in research reports. This is perfectly reasonable when considering that the goal with such projects is often to develop highly accurate and efficient algorithms, e.g. for participating in competitions where performance on a specific shared test dataset is the main assessment criterion.

In summary, there is an unsatisfied need among researchers and practitioners for being able to extract entities beyond the classic set of named entities from text data in an efficient and predictably accurate fashion for the purpose of constructing multi-mode network data that support users in answering substantive question about socio-technical networks (Barthelemy et al., 2005; Parastatidis et al., 2009; C. Roth, 2006). This thesis addresses this need in two ways: First, by devolving a computational solution to this problem (this chapter). Second, by demonstrating the transition from learned models to an end-user technology and the application of these models and technology to large-scale network data (next chapter).

3.2 Goal Definition, Requirement Specification and Strategies for Achieving Objectives

The goal and deliverable for the project described in this chapter is an entity extractor that end-users can apply as part of the process of constructing multi-mode, socio-technical network data from texts. To provide end-users with this technology, I add an inference mechanism that uses the prediction models designed, learned and evaluated as described in this chapter to AutoMap software, where this new functionality is expected to improve the status quo of entity extraction. The extracted entities can then be used to conduct classic content analysis or to construct network data, which can be further analyzed with tools such as ORA. The ORA software is tuned for the kind of network data and ontological text coding that AutoMap supports (Carley et al., 2007; Carley, Reminga, Storrick, & Columbus, 2011).

From an NLP perspective, the research question that typically drives the development of entity extractors is usually formulated like this: How can we build or improve an entity extraction algorithm or system that leads to the comparatively most accurate results? Points of comparison are often a baseline and/or the best-performing alternative solutions. In this thesis, I shift the focus from further gains in accuracy to gains in the practical usefulness of the extracted data for conducting network analysis. Thus, my research question for this chapter is this:

> *Research question:*
> *How can we build an entity extractor as part of a relation extraction system that supports users in analyzing networks and addressing substantive questions about socio-technical networks?*

This question will be further specified in this chapter. From a network analysis perspective, this question has to be answered before the aforementioned NLP-oriented question becomes applicable. It is important to highlight that this research question does not contradict with the one typically asked in NLP; both questions are critical. Rather, my question complements the one from the NLP perspective because accuracy is one among multiple important criteria for entity extraction; yet other criteria include the appropriateness of the coding schema and methods for analyzing the resulting data (Schrodt, 2001). In the next section, I formalize my research question: I describe how entity extraction and node linkage are currently handled in AutoMap (3.2.1), then define the requirements for a new entity extractor (3.2.2), and develop a solution per requirement (3.2.3 to 3.2.6).

3.2.1 Status Quo of Entity Extraction in AutoMap

AutoMap is a text mining tool that provides routines for information extraction and relation extraction (for a detailed description of AutoMap see Carley et al., 2007; Diesner & Carley, 2004). In AutoMap, concept networks are called semantic networks, and multi-mode networks are called meta-networks (Carley, Columbus, et al., 2011). The method used for coding text as networks in AutoMap was originally called "map analysis" (Carley, 1993); a reflection of its purpose to extract mental models of individuals and teams from texts (Carley, 1997a; Carley & Palmquist, 1991). Later, the method was referred to more generally as "network text analysis" (NTA), which basically works as follows (Carley, 1997b; Popping, 2003): the user creates a dictionary or thesaurus, which associates terms as they occur in the text data with user-defined concepts that represent variables of interest. The software assists the user in this process, e.g. by suggesting a set of relevant terms according to (weighted) term frequencies. Concepts represent pieces of information that are necessary for answering a research question; similar to codes in qualitative text analysis (Bernard & Ryan, 1998). The software then applies the thesaurus to the text data by translating any matching terms into the respective concepts. Finally, the concepts are linked into edges by using windowing; a proximity-based approach (Danowski, 1993). The main assumption with map analysis and NTA is that these methods support the extraction of meaning from texts by finding or establishing links between concepts and conducting network analysis of the resulting data (Carley, 1994, 1997b; Mohr, 1998; Monge & Contractor, 2003; Popping, 2003; Van Atteveldt, 2008). Entity extraction and linkage in AutoMap are computer-assisted processes. This means that the software applies a set of text pre-processing and link formation rules, which together form the so called "coding scheme". The coding scheme is defined by humans (G. W. Ryan & Bernard, 2000). Section 5.2.2.1 provides more details on the steps needed for text coding in AutoMap.

In summary, the key piece needed not for only entity extraction, but also for text coding in general in AutoMap is a thesaurus. Section 5.2.2.1.1 reports in detail on preparing a thesaurus. For generating concept networks, a thesaurus needs to contain two columns: text terms on one side, and the associated concepts on the other side. For creating multi-mode network, an additional column is needed that associates concepts with entity classes. In AutoMap, concepts and entity classes can have attributes, but the (types of) attributes are neither predefined nor required. Similar to the creation of code books for content analysis, creating thesauri is a very time-consuming and cumbersome process, even if it is computer-supported, and requires people specifically trained for this task (Corman et al., 2002; King & Lowe, 2003; Krippendorff, 2004; Schrodt et al., 2008). Typically, thesauri are validated by assessing the degree to which one person assigns the same code to the same text over time (intra-coder reliability) or to which multiple people assign the same code to the same text (inter-coder reliability). We have been adding a plethora of features to AutoMap to make the thesaurus construction process more efficient, such as generating lists of salient terms and N-grams based on their (weighted) frequencies, and stemming terms into their morphemes, which potentially allows for more hits per term (Diesner & Carley, 2004, 2008a).

3.2.2 Requirements for Entity Extractor

We identified a set of seven criteria that are important for an entity extractor that serves the purpose stated for this project in general and in AutoMap specifically. In order to find these criteria, I began by specifying what type of network analysis the extracted entities data should support in the end. Different approaches to network analysis are suited for different purposes, and can be placed on a spectrum ranging from social network analysis to network science. Table 50 summarizes key characteristics of these poles as they are relevant for this section, and provides examples of typical applications.

Table 50: Characteristics of Network Analysis approaches

	Network Science	Social Network Analysis
Goal	- Identify, formally describe, model, and test hypothesis and advance theories about properties, dynamics and evolution of graphs, link data, and relational data.	- Answers substantive questions and advance theories about the individual and collective behavior and cognition of social agents. - Develop and test hypothesis and theories about implications and causes of the properties, dynamics and evolution of network data.
Research process (Figure 2)	- Focus on the computational analysis of data w.r.t. to a research question. Existing or benchmark datasets are often used.	- Data collection is often part of the analysis process.
Scalability	- Focus on large-scale graphs and change of graph properties as network sizes change.	- Traditionally, datasets, methods and tools were focus on network data of small to moderate size. This has shifted to ambitions to test and develop theories about networks of any size.
Exemplary application domains	- Technical infrastructures such as telecommunication networks and the internet (Barabási & Albert, 1999; Eagle & Pentland, 2006). - Other sizable socio-technical networks, e.g. geopolitical entities (Auerbach, 1913; Newman, Strogatz, & Watts, 2001; Simon, 1955). - Online social networks and social media data (Adamic & Huberman, 1999; Leskovec et al., 2007).	- In social sciences and organization science, mainly: - Innovation diffusion (Coleman, Katz, & Menzel, 1966; Kraut, Rice, Cool, & Fish, 1998) - Group structure and processes (Milgram, 1967; Sampson, 1968) - Communication networks (Monge & Contractor, 2003) - Learning and information processing of social agents (Carley & Palmquist, 1991; Collins & Loftus, 1975)

Ultimately, the goal with this project is to provide a technology that combines the advantages from both sides of the spectrum shown in Table 50. This means that I aim for a solution that extracts data which allows users to gain deep and rich knowledge about network of any size, to formally describe this knowledge, and to answer substantive questions about networks (Corman et al., 2002; Hirst, 2006). I broke this high-level goal down into separate, more specific goals that are detailed in Table 51. These goals are relevant for this thesis, but are not a comprehensive list of requirement for network data collection tools.

Table 51: Goals for entity extractor

Goal	What does that mean?	Why is it relevant?	How does it improve the status quo of AutoMap?
1. Automation	The ability to automatically collect one-mode and multi-mode network data.	Contributes to scalability. Reduces time and labor costs. (Corman et al., 2002)	Extracting networks in AutoMap requires the semi-automated construction and/or adaption of thesauri. This is very time-consuming and laborious (see section 5.2.2.1.1 for a description of thesaurus preparation).
2. Abstraction of terms to concepts or higher level aggregates	The ability to associate terms with higher level abstractions, e.g. concepts. In Entity Extraction, the entity classes are higher level aggregates.	Enables analyses on different levels of granularity and aggregation. (Monge & Contractor, 2003)	The data structures used for network representation in AutoMap and ORA supports the association of terms with concepts (and attributes of) certain entity classes. Being able to efficiently extract these associations in AutoMap creates a more capable and efficient tool chain.
3. Generalization	The ability to identify new and unseen instances of entity classes and entity attributes.	Contributes to greater flexibility in extracting network data from new corpora. Reduces time and labor costs.	AutoMap is constrained to only find entities that are specified in a thesaurus. In order to also find and classify new terms, the thesaurus needs to be extended in a time-consuming, semi-automated way (see section 5.2.2.1.1 for details).
4. Support end-users in addressing substantive and meaning-ful questions about socio-technical networks	Being able to go from texts to network data to knowledge. Provide publicly available entity extractor that is readily useable.	Contributes to practical usefulness of network analysis. Allows for answering substantive questions about networks. (Alderson, 2008; Krackhardt & Carley, 1998)	ORA already supports the automated analysis of large-scale, multi-mode network data. Being able to efficiently extract this data with AutoMap creates a more capable and efficient tool chain.
5. N-gram detection	Correctly locate the boundaries of unigrams and multi-word entities.	Default requirement for NER. (Ratinov & Roth, 2009)	AutoMap provides a probabilistic solution for extracting unigrams only (Diesner & Carley, 2008a).
6. Allow terms to	The same term can belong to	Contributes to the disambiguation of	AutoMap can assign one term to one concept only, and one concept

belong to multiple entity classes instead of just one	multiple entity classes given a term's meaning and context. Such terms can be homonyms or identical terms.	homonyms. Prevents the loss of relevant information.	to one meta-network category only. This goal addresses the first step.
7. Entity Extraction (as opposed to focus on Named Entity Extraction)	Extract entities that are referred to by a name or not, which is particularly relevant for entity classes where many are generic identifies, such as references to roles and collectives.	Contributes to answering substantive questions about socio-technical networks, e.g. about culture and ethnography. (Diesner & Carley, 2008a)	ORA supports the automated analysis of unnamed and unnamed entities. Being able to efficiently extract these entities with AutoMap creates a more capable and efficient tool chain.

3.2.3 Review and Selection of Method to Enable Automation, Abstraction, and Generalization

Achieving automation, abstraction and generalization (goals 1-3) requires the selection of an appropriate REX method while keeping the subsequent use of entities for network construction in mind. The solution developed herein satisfies these three requirements by choosing a method that best covers the stated goals: this method selection is based on my review of the main families of methods that are available for generating concept networks from text data as summarized in Table 52. Note that the focus with Table 52 is on methods for generating word networks, not methods for analyzing them. A more detailed review of these methods is provided in Diesner and Carley (Diesner & Carley, 2010c), and a review of current computational methods in Mihalca and Radev (2011). Some of the listed methods are outdated and hardly used anymore, but have laid the foundations for further advances. The semantic web, for instance, can be considered an extension of definitional semantic networks. Furthermore, some of the seminal methods overlap. Map analysis, for example, borrows elements from spreading activation theory and knowledge representation in artificial intelligence. Also, most of the listed methods were not developed with the goal of providing input to network analysis or to handle just the extraction of entities and relations, but rather for transforming texts into relational presentations of information and knowledge in order to solve problems in specific application domains. I include those methods in this review not only to be comprehensive, but also to show that the construction of concept networks has roots in many disciplines. This review suggests that machine learning methods that are based on probabilistic graphical models (PGM) (group 17) fulfill the requirements of automation, abstraction and generalization. Therefore, I selected this methodological approach for this project. The selection of a specific PGM-based method is described in section 3.3. However, this choice implies one limitation: in order to reason about the meaning of the extracted data, further network analysis is needed once the data have been constructed. This task is addressed in the next section.

Table 52: Review of family of methods for generating word networks

Families of methods for constructing word networks and seminal papers	**Automation** No: manual Yes: automated CoSu: computer supported	**Abstraction** No: use terms verbatim Yes: map terms to higher level representation	**Generali-zation** No: deterministic Yes: find new instances	**Steps needed to reason about meaning of network data**
1. Discourse Representation Theory (Kamp, 1981)	No	Yes	No	Data construction process
2. Mind maps (Buzan, 1974)	No, CoSu	Yes	No	Data construction process Data analysis
3. Concept maps (Novak & Gowin, 1984)	No, CoSu	Yes	No	Data construction process Data analysis
4. Hypertext (Trigg & Weiser, 1986)	CoSu	Yes	No	Network analysis Inference
5. Qualitative text coding according to Grounded Theory (Glaser & Strauss, 1967; T. Richards, 2002)	No, CoSu	Yes	No	Data construction process Data analysis
6. Mental Models according to Spreading Activation (Collins & Loftus, 1975; Collins & Quillian, 1969)	CoSu	No	No	Data analysis
7. Knowledge representation in artificial intelligence, assertional semantic networks (Shapiro, 1971; Woods, 1975)	Yes	No	No	Inference

8. Definitional semantic networks incl. networks built by using an ontology (Berners-Lee et al., 2001; Fellbaum, 1998)	Generation: no Usage: yes	Yes	No	Data analysis Inference
9. Semantic Web (Berners-Lee et al., 2001; Van Atteveldt, 2008)	Generation: no Usage: yes	Yes	No	Information retrieval
10. Case Grammar and Frame Semantics (Fillmore, 1968, 1982)	Generation: no Usage: yes	No	No	Data analysis
11. Frames (Minsky, 1974)	Generation: no Usage: yes	Yes	No	Data analysis
12. Semantic Grammars (Franzosi, 1989; C. W. Roberts, 1997a)	CoSu	Yes	No	Data analysis Statistical analysis
13. Semantic network in communication science (Danowski, 1993; Doerfel, 1998; van Cuilenburg, Kleinnijenhuis, & de Ridder, 1986)	CoSu, Yes	Yes	No	Network analysis
14. Centering Resonance Analysis (Corman et al., 2002)	Yes	No	No	Network analysis
15. Map Analysis, Network Text Analysis in Social Science (Carley & Kaufer, 1993; Carley & Palmquist, 1991)	CoSu	Yes	No	Network snalysis
16. Event Coding in political science (King & Lowe, 2003; Schrodt et al., 2008)	CoSu	Yes	No	Statistical analysis
17. Machine learning based on probabilistic graphical models (Howard, 1989; Pearl, 1988)	Generation: no (orig.) to yes Usage: yes	Yes	Yes	Inference Network analysis

3.2.4 Review and Selection of Approach to Support Addressing of Substantive and Meaningful Questions about Socio-Technical Networks

The fourth requirement for the entity extractor is the generation of data that allows for addressing substantive questions and reasoning about the meaning of networks. What does it mean for network data to support meaningful analysis? In this section, I discuss this question and conclude with the selection of an approach. The meaning of relational representations of language and knowledge has been extensively discussed in the linguistics and artificial intelligence literature (Hirst, 2006; Ogden & Richards, 1923; Woods, 1975). There, concept networks that represent meaning are called semantic networks (for a brief synopsis see Diesner & Carley, 2011a; Sowa, 1992; Woods, 1975). A unifying assumption across various approaches to semantic networks is that the meaning of concepts can be inferred from a concept's context as explicitly or implicitly provided in text data or the network data (Collins & Quillian, 1969; Griffiths et al., 2007; Minsky, 1974; Shapiro, 1971; Weaver & Shannon, 1949). According to Hirst (2006), further progress in extracting meaning from texts will require a combined consideration of subjective authorial intent, subjective interpretations of the reader, and the extraction of objective representations of meaning from large-scale corpora. In the network analysis literature, the meaning of word networks has been hardly discussed. There, the assumption is that a node's meaning results from its context and network position; both of which can be described via network analytical measures (Carley, 1997b; Carley & Kaufer, 1993; Carley & Palmquist, 1991; Doerfel, 1998; Mohr, 1998). Context here means the structural environment of a node, typically starting from the ego-network. Detecting a node's meaning basically requires completing the network analysis process as outlined in Figure 2. However, there is no guarantee that a concept network or its analysis will be meaningful. Moreover, it is easy to read patterns and meaning into networks, for example by making heuristic use of network visualizations (Bernard & Ryan, 1998).

A synthesis of prior work on enabling the reasoning about the meaning of word networks is provided in the last column of Table 52; suggesting that there are five options for achieving this goal: (1) some methods require humans to go through a cognitive, typically manual or computer-supported, process of creating concept networks. This data construction process requires the representation of the meaning of concepts and relations as perceived by the people creating the data. With some of these methods, meaning can also be obtained by interpreting the resulting data. For example, when applying grounded theory methodology to construct structural models based on text data, the resulting data are assumed to be inherently meaningful, but require the analysts' interpretation with respect to their research question (Glaser & Strauss, 1967). In general, three types of analysis can be employed to get to the meaning of the relational data: (2)

statistical analysis, (3) network analysis, and other types of (4) data analysis such as qualitative interpretations. Note that not all methods with which concept networks are generated assume the usage of network analysis methods to reason about the resulting data. For example, semantic web data are generated to support information retrieval, and relational data generated with event data coding methods in political science are typically analyzed with non-relational, but statistical methods. Finally, some methods involve the possibility of conducting (5) inference on the generated data.

There are two more strategies for supporting the construction of meaningful data; both of which are an integral part of many of the outlined methods and cut across the five strategies outlined above: First, concept networks can be constructed by using structured variables that are motivated by theory (Corman et al., 2002; Van Atteveldt, 2008). Second, meaningful concept networks (in the definitional sense of "semantic networks") can be generated by applying predefined classification schemata, i.e. specifications of the set of possible elements (denoted in ontologies) and relations between them (fixed in taxonomies) in a given domain (Berners-Lee et al., 2001; Gerner et al., 1994).

In order to ensure that the entity extractor developed herein supports the construction of network data that allows for meaningful analysis, I combine the following elements which are all selected from the options discussed above:

1. Use an *ontology* that is grounded in theory from the social sciences and defines the entity classes that are typically relevant for socio-technical network (section 3.2.5).
2. Use *probabilistic graphical models* as the method for generating a prediction model that retrieves instances of these entity classes from text data (section 3.3).
3. Generate concept networks that are structured such that all entity classes, links between entities, and attributes of nodes and entities can be analyzed via *network analysis*, *statistical analysis* and *visualizations* with an existing toolkit (ORA: Carley, Reminga, et al., 2011) This is demonstrated in chapter 5.

3.2.5 Selection of Ontology

The standard set of entity classes for Named Entity Recognition in NLP comprises *agents*, *organizations*, *locations* and miscellaneous *other* entities. In political science, the categories considered for event coding are *agents* and *events*, and for both of these categories, elaborated sets of subtypes exist, which are continuously updated in a collaborative fashion (Schrodt et al., 2008). In organization science, Krackhardt and Carley (1998) have developed a multi-mode and multi-plex model called PCANS that defines the set of relevant entity classes; namely *agents*, *tasks* and *resources*. PCANS also specifies primitives or general templates for the possible relation between these classes. These primitives result from the logical and temporal ordering of activities, and can be represented as combinations of matrices of the considered entity types. Carley (2002a) has extended the PCANS model into the meta-matrix model in two ways: first, she extended the set of categories to represent the *who* (agent, organizations), *what* (task, event), *when* (time), *where* (location), *why* (emotions, beliefs) and *how* (resources, knowledge) of events. Second, she developed a plethora of network analytical measures that are defined over these nodes types and their combinations. These measures are implemented in ORA (Carley, Reminga, et al., 2011). In general, most network analytical measures are defined independently of node types (Wasserman & Faust, 1994). Thus, these measures are assumed to be appropriate for analyzing networks of any type, including social networks and generic graphs. Tailoring measures to certain entity classes and types of graphs as supported with the meta-matrix model and by ORA allows for more detailed and richer network analysis. The meta-matrix model has been previously tested, applied and validated in a variety of contexts such as situational awareness in remote work teams (Weil et al., 2008), collaboration in groups (Cataldo et al., 2006), public health (Merrill, Bakken, Rockoff, Gebbie, & Carley, 2007), and geopolitical groups (Carley et al., 2007). The definition of entity classes, attributes, subtypes of classes, and respective measures for the meta-matrix keeps being adjusted and updated. In summary, I chose to use the meta-matrix model as an ontology for defining the entity classes that the entity extractor needs to recognize. This choice enables the collection of rich network data for which measures have already been defined and validated, and for which an analysis tool is readily available.

3.2.6 Selection of Solutions to Entity Extraction, N-gram Detection, and Non-Exclusive Term Classification

Entity Extraction: The meta-matrix model comprises various classes where entities are often not referred to by a name, such as tasks and resources. In the next step, training data needs to be selected that contain examples of a mix of named and unnamed entities for the entity classes of interest. The selection of an appropriate learning dataset is presented in section 3.3.1.

N-gram Detection: Each instance of a relevant entity class needs to be detected from its beginning to its end, whether it's a unigram or a multi-word expression. This is a token labeling task (Sarawagi, 2008), which I herein refer to as "boundary detection". In fact, with entity extraction via machine learning, a boundary class label is predicted for every token in the text data, but while only those tokens that do not fall into the "outside" boundary class are output to the user, the boundary label for every token counts for accuracy assessment. In prior work, various classification schemas for boundary classes have been used: the simplest one is the BIO (begin, inside, outside) schema, more advanced is BIEO (begin, inside, end, other), and even more detailed is BIEOU (begin, inside, end, other, unigram) (Ratinov & Roth, 2009; Sarawagi, 2008). Choosing a boundary class model means making a tradeoff between expressiveness versus keeping the number of parameters for learning small. A model for a given project can be chosen by testing the performance of various models on data, or by building upon prior empirical results. I chose the latter approach: Ratinov and Roth (2009) showed that BIEOU outperforms BIO by 0.5% to 1.3% on two training datasets, respectively. These datasets are similar in their genre and entity classes to the data used for learning herein. Currently, the entity extraction feature in AutoMap that was built by using a machine learning approach based on probabilistic graphical models is only capable of locating and classifying unigrams, regardless of whether they are constituents of N-grams or not (Diesner & Carley, 2008a). Adding a routine that properly handles the detection of multi-word expressions will help to improve the extraction of concept networks as well as meta-networks. Since concept networks are one-mode networks, the only applicable entity extraction task for these networks is boundary detection.

Allow terms to belong to multiple entity classes instead of just one: From a practical and realistic perspective, entity extraction is a non-exhaustive, non-exclusive process. This means that not all words are relevant entities, but those that are relevant might fall into multiple categories depending on the terms' identity and context. What does that imply for the selection of a machine learning method? Since in fact most words in a text do not belong to one of the meta-network categories, the prediction model needs to be able to handle very sparse data. Sparse here means that most terms fall into the "O" (outside) category of the boundary coding schema. Thus, the method must not strongly rely on transition probabilities between entity classes, but needs to exploit other information from the text data. Frequently used alternative clues are characteristics of the terms themselves, long-distance information from sequential data, and the relationship between a term and its label (McCallum, 2005; Sarawagi, 2008). Currently, the way thesauri are processed in AutoMap requires that each term is mapped to only one concept, and each concept to only one meta-network category. Thus, thesauri in AutoMap are currently structured this way. Outputting thesauri where the same terms can be mapped to multiple entity classes will enable the disambiguation of homonyms and identical terms that belong to different categories in different situations. In order to correctly map these entries to their respective occurrences in the text data, further features per word need to be considered, such as parts of speech or local context. Considering this modification to thesauri for actual text coding projects will require changes to the AutoMap backend that are not subject of the work for this thesis, but the outcome of this thesis is a precondition for this next step.

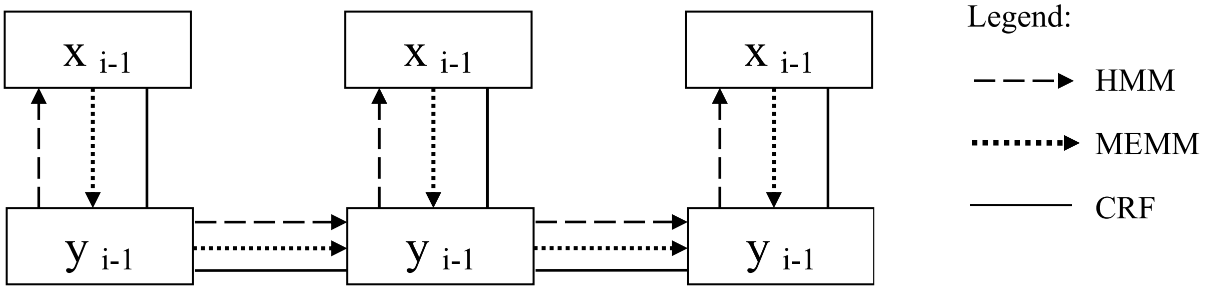
3.3 Method

Summarizing the findings from the presented requirements analysis, the following criteria are appropriate and necessary for an entity extraction method:

- A machine learning technique based on probabilistic graphical models (PGM).
- A technique that can handle the sparse distribution of relevant entities across text data.
- A technique that allows for assigning tokens with identical surface forms to different categories. .
- A technique that is able to exploit long distance information in sequential data. Sequential here means that when generating text data, terms and class labels are not drawn independently from some distribution, and terms and labels may show sequential correlations. Due to the sequential nature of unstructured text data, a PGM is needed that is able to capture and exploit dependencies of tokens and labels (Sarawagi, 2008).

Given the availability of suitable training data for the task at hand as will be described in section 3.3.1, I chose to use a supervised learning approach. In general, sequential supervised learning makes probabilistic predictions about the relationship between consecutive tokens *x* and a *y* label for every token (Dietterich, 2002). For this project, each token is an *x*, and the respective class label is the *y*. The learning goal for this project can be formulated as follows: Learn a prediction model or classifier *h* that for each sequence of *(x,y)* suggests an entity sequence *y=h(x)* that generalizes with predictable accuracy to new and unseen data. Several PGMs for sequential learning satisfy the identified requirements. I briefly describe eligible models along the dimensions of directionality and the type of distribution they estimate, because these two characteristics are relevant for the given task. Figure 9 shows a schematic depiction of the PGMs discussed in this section.

Figure 9: Graph structure of selected Probabilistic Graphical Models



The directionality of arcs in the model represents assumed logical dependencies. In directed PGMs, every node is conditioned on its parent(s). In undirected models, distributions are factored into local likelihood functions for each clique of variables. PGMs can be divided into generative models and conditional models, aka discriminative models:

With generative models, a joint distribution of the form *P(x,y)* is estimated. An example for generative models that are frequently used for entity extraction are Hidden Markov Models (HMM). An early system that successfully used HMM for NER is IdentiFinder (Bikel et al., 1999) , which exploits multiple features of words and achieves an NER prediction accuracy of F= 94.9%.

Conditional models estimate a conditional distribution of the form *P(y|x)*. For the given task, the output generated from conditional models, i.e. the most likely class label sequence *y* per token sequence *x*, is what we are truly interested in, while explaining how the token sequence was generated from the class labels through an assumed probabilistic (generative) process is irrelevant. A highly accurately performing conditional PGM for NER are Conditional Random Fields (CRF) (Lafferty, McCallum, & Pereira, 2001; Sha & Pereira, 2003). CRF have shown to outperform alternative generative models. For instance, Lafferty et al. (2001) obtained an error rate of 5.55% with CRF, 6.37% with Maximum Entropy Markov Models (MEMM), and 5.69% with HMM. MEMM are another discriminative model (Borthwick, 1999).

In general, the accuracy rates obtained with HMM are comparable to those achieved with conditional models. The main disadvantage of HMM are their strictly local properties: HMM lack the ability to directly pass information between non-adjacent *y* values. Instead, such information must pass through the intervening *y*'s (Dietterich, 2002). Also, each token is assumed to be generated from the corresponding class label only. Thus, information about other nearby labels cannot be considered. However, information about not directly co-located elements is particularly valuable when working with sparse data and for multi-word units that are longer than two tokens. Conditional models do not have this limitation; they allow for considering arbitrary features of *x*, including global and long-distance features (Dietterich 2002).

Within the group of conditional models, MEMM have led to higher error rates than generative models (Lafferty et al., 2001). This limitation been explained with the "label bias problem": MEMM are a log-linear model that maximizes the conditional probability of each label *y* given the previous label y_{i-1} and the current token x_i. Once this maximization is done, MEMM use maximum entropy to compute the highest conditional likelihood of all *x*: $\prod P(y_i| x_i)$. The label bias occurs in the first step of this process: each y_{i-1} has to pass all of its probability mass to the adjacent label y_{i-1}, even if a token x_i hardly fits this choice (Lafferty, McCallum and Pereira 2001). Since CRF do not have the same local constraint, they can delay this decision until a good fit has been found.

CRF feature some additional advantages: First, they can find global optima in sequential data with respect to the target function specified for this project. Second, CRF can take arbitrarily large numbers of features into account. In fact, since the identity of every word can be used as a feature, the number of feature can easily be in the tens of thousands. This largely exceeds the handful of features typically used with more local models. Therefore, more of the information available in text data can be exploited, including weak contributors, which are crucial for working with sparse data. Third, CRF allow for considering long-distance information between at least the tokens.

The main caveat with CRF is that they require high time costs for training. This is mainly due to performing global search with a reasonably sized gradient in a large feature space. However, once the model is learned, inference time is not subject to this constraint. Therefore, applying the model in end-user applications is fast and scalable.

In summary, given the outlined characteristics, strengths, weaknesses and empirical accuracy rates for the discussed PGMs, I chose CRF as the PGM based machine learning technique for this project. This choice is supported by prior work: Sarawagi (2008) concludes that for data at the level of heterogeneity that we aim to provide an entity extractor for, i.e. mainly unstructured data from well-defined genres and domains, conditional model and learning based on enough training data are the state of the art approach to this task. In our case, the domains to be covered are news data and other reports of interactions and events in organizations.

In contrast to HMM and MEMM, CRF model the relationship among each label y_i and its predecessor y_{i-1} as a Markov Random Field (MRF). MRF are an undirected PGM that is conditioned on x only. In CRF, the distribution $P(y|x)$ is computed as a normalized product of potential functions M_i, which are computed as shown in Equation 4 (Lafferty et al., 2001; Sha & Pereira, 2003):

Equation 4

$$M_i(y_{i-1}, y_i \mid x) = (\exp\left(\sum_{\alpha} \lambda_\alpha f_\alpha(y_{i-1}, y_i, x) + \sum_{\beta} \mu_\beta g_\beta(y_i, x)\right)$$

In Equation 4, the f_α expression is an edge feature that represents the transitions between labels and tokens. Furthermore, g_β is a vertex feature that represents the emission of an entity from a term sequence. Feature vectors f_α and g_β are fixed, boolean vectors. Most of the time, a feature will be switched off or be zero (sparse data), and is turned on only when applicable. For example, the word identity feature, which this implementation includes, is only switched on when *x* contains that particular term. When a feature is switched on, the specific learned weight per feature, i.e. λ_α and μ_β, becomes applicable.

In order to normalize the scores of the potential functions, the M_i are typically multiplied with *1/Z(x)*. Here, *Z* is a normalizing constant parameterized on the sequence *x*. Finally, the conditional probability of the entire label sequence *P(y|x)* is computed as shown in Equation 5. Note that in Equation 5, both *y* and *x* are arbitrarily long vectors.

Equation 5

$$p_\theta(y \mid x) = \frac{\prod_{t=1}^{n+1} M_i(y_{i-1}, y_i \mid x)}{\prod_{i=1}^{n+1} M_i(x)_{start,stop}}$$

3.3.1 Learning Data

Supervised machine learning requires marked up or labeled data for training and testing. Since the goal here is to predict a boundary and category for every entity, a dataset is needed where the start, end and category of all relevant entities are marked up. Building a high quality learning dataset is expensive because it requires humans for this task, a sufficiently high rate of intercoder reliability, and a sufficiently large number of marked up examples. No such dataset that covers instances of the meta-network categories has yet been created in our group. Therefore, I had to defer to external sources. In order to find the most suitable training dataset for this task, I reviewed the major datasets that are available to researchers for information extraction purposes. Table 5 provides a reference and a short overview of the main characteristics of these datasets. Some of these datasets cover the main set of entity classes that are typically considered in information extraction, but no further subtypes. These datasets are shown in Table 53, which also specifies how these main categories are referred to in the meta-network model.

Table 53: Entity class review I: Models and datasets without subtypes

Entity class	Meta-network	ACE-2, TIDES	NYT	CoNLL-2003
Person	x (Agent)	x	x	x
Organization	x	x	x	x
Location	x	x	x	x
Facility	x (Location)	x		
GPE	x (Location)	x		

In some of these datasets, specific and generic instances of categories are not distinguished from each other. This would be problematic for the types of analysis we aim to support: in our practical work we often have seen that when identifying key agents in networks, generic nodes such as "president" often rank very high because they subsume references to multiple individuals, but are not as meaningful as the name of a specific president (Diesner & Carley, 2005b). This problem applies to references to roles of people and organizations in general. Therefore, datasets that allow for distinguishing between generic and specific entities are needed. The applicable datasets are compared in Table 54, which covers the same entity classes as Table 53. In addition to that, Table 54 lists the available subtypes per entity class and lines them up across corpora where possible.

Table 54: Entity class review II: Models and datasets with subtypes

Entity class	**MUC6, 7 (NE task)**	**Subtypes (IE task)**	**ACE 2004, 2005**	**Subtypes**	**BBN**	**Subtypes**
Person	x	name	x	individual ('05)	x	(name, desc)
		alias		group ('05		
		title		indefinite ('05)		
		types (7): other, military, civilian				
Org.	x	name	x	government	x	government (name, desc)
		alias		commercial		corporation (name, desc)
		descriptor		educational		educational (name, desc)
		type:		non-profit ('04)		political (name, desc)
		government,		non-governmental ('05)		
		company, other		religious ('05)		religious (name, desc)
				media ('05)		hotel (name, desc)
				entertainment ('05)		hospital (name, desc)
				medical-science ('05)		museum (name, desc)
				sports ('05)		
				other ('04)		other (name, desc)
Location	x	city	x	address	x	
		province		boundary		border (name)
		country		celestial		
		region		water body		lake sea ocean (name)
		unknown		land region natural		river (name)
		water (7)		region local ('04)		region (name)
		airport (7)		region sub-nat. ('04)		
				region national ('04)		
				region general ('05)		
				region international		continent (name)
				other ('04)		other (name)
Facility			x	airport ('05)	x	airport (name, desc)
				plant		
				building ('04)		building (name, desc)
				bldg. on grounds ('05)		

			sub area building ('04)		
			sub area facility ('05)		
			bounded area ('04)		
			conduit ('04)		bridge (name, desc)
			path		highway street (name, desc)
			barrier ('04)		attraction (name, desc)
			other ('04)		other (name, desc)
GPE		x	continent	x	
			nation		country (name, desc)
			state or provine		state province (name,
			county or district		desc.)
			city or town ('04)		
			population center ('05)		
			GPE cluster ('05)		city (name, desc)
			special ('05), other		other (name, desc)

The datasets considered in Table 54 go beyond the standard set of entity classes by providing markups for additional classes and their subtypes as shown in Table 55. The point of reference in Table 55 (leftmost column) is the set of categories defined for the meta-network model. This comparison shows that no dataset covers all of the meta-network categories, but the BBN dataset comes closest to that by covering all but the "beliefs" category. However, in BBN, one subtype of agents and organizations is "religious", which captures the notion of agents adhering to a belief. This label approximates the purpose behind the belief class in the meta-network.

Table 55: Entity class review III: Additional entity types

Meta-network entity class	**MUC6, MUC7 (NE task)**	**Subtypes (IE task)**	**ACE 2004, ACE 2005 (*= value of entry)**	**Subtypes**	**BBN**	**Subtypes**
Resource	Artifact	ID,	Vehicle	air, land, water,	Product	weapon (name, desc)
	(IE	description		subarea vehicle,		vehicle (name, desc)
	task)	type (7): air,		other ('04),		other (name, desc)
		ground, water		underspec. ('05)	Substance	food, drug, nuclear,
			Weapon	blunt, exploding,		chemical, other
				sharp, chemical,	Plant	
				biological,	Animal	
				Nuclear,	Disease	
				other ('04),		
				underspec. ('05)		
	Money		Money ('05)*		Money	
Time	Time	7: descriptor,	Time ('05)*	TIMEX2, incl.:	Time	
		start, end		present, past,	Date	date, duration,
		type: before,		future		age, other
		on, after,		type: within,		
		between		start, end, as of,		

	Date		before, after		
Knowledge				Law Language Work of art	(name) (name) book (name) play (name) song (name) painting (name) other (name)
				NORP	nationality (name) religion (name) political (name) other (name)
		Contact ('05)*	email, phone#, URL	Contact info	address, phone #, other
Belief					
Attributes	Percent	Percent ('05)*		Percent Ordinal Cardinal Quantity	 1D, 2D, 3D, energy, speed, temperature, weight, other

Table 56 compares the additional attributes or classifications that the reviewed datasets contain per entity class. In BBN, the generic versus specific distinction as well as further subtypes of entity classes (if applicable) are directly encoded in the category label itself, while in MUC and ACE, any additional information is marked up as separate attributes per entity. In general, BBN integrates features from different datasets: similar to ACE, it annotates numerous subtypes of entities, and like in MUC, BBN separates all entities into named entities, temporal expressions and numerical expressions.

Table 56: Entity class Review IV: Additional attributes for entities

Meta-network	**MUC6, MUC7 (NE task)**	**ACE 2004, ACE 2005**	**BBN**	**ACE-2, TIDES**
For Per, Org, Loc:	For each entity:			
specific generic	named entity	name nominal pronoun	named entity	name nominal pronoun
		for each entity (2nd attribute):		for each entity (2nd attribute):
	temporal expression number expression	negatively quantified non-ref./attribut./ascriptive specific referential generic referential under-specified referential	temporal expression number expression	generic specific

The only entity class that is treated differently in the discussed learning datasets than in the meta-network model is the activities category: in the meta-network model, instances of the “task” and “events” class comprise a single word or a short phrase, such as “participate in”. Nodes of these types can be linked to nodes from the same or any other entity class. A similar approach to event coding is typically taken in political science, where events are terms that can have a valence value and take agents as their arguments (Gerner et al., 1994; Goldstein, 1992; King & Lowe, 2003; Schrodt et al., 2008). In that domain, the types of events and agents are predefined, while specific instances of these entity classes are identified from the actual text data via shallow parsing techniques. The goal with this type of event coding is to identify who does what to whom. In contrast to that, in NLP-style information extraction, event coding is conceptualized as a slot filling or relation extraction task: an event or scenario consists of various entities of predefined types that play certain, predefined roles or have certain relationships with each other. These events are typically very specific and cannot be expected to generalize well to other types of activities. Table 57 compares the event coding approaches in the potential learning datasets. This comparison shows that the ACE 2005 data encodes a variety of events that are relevant for asking substantive questions about socio-technical networks. Moreover, ACE 2005 offers predefined valence values (polarity) for these events. BBN lacks these features, but offers a different advantage: event mark-ups in BBN are closest to the way that the meta-network model represents activities. However, the types of events considered in BBN are confined to specific mentions of wars, hurricanes and other events as well as games, such as sports games.

Table 57: Event coding review

Meta Net-	MUC6, MUC7 (NE task)	ACE 2005	Subtypes	BBN	Subtypes
Event	6: management succession:	life	be born, mary, divorce, injure, die	Event	war (name)
	succession	movement	transport		hurricane (name)
	in and out	transaction	transfer ownership		other (name)
Task			transfer money	Game	
		business	start org, merge org, end org, declare bankruptcy		
	7: air vehicle launches:				
	launch event	conflict	attack, demonstrate		
	vehicle info	contact	meet, phone, write		
	payload info	personnel	start position, end position, nominate, elect		
		justice	arrest jail, release parole, trial hearing, charge indict, sue, convict, sentence, fine, execute, acquit, appeal, pardon		
		arguments:	who, when, where, instrument, price, target		

		values: crime, sentence, job title Per event: polarity (occurred or not) tense (past, presence, future) genericity (generic, specific) modality (asserted, other)	

In summary, the review of potential learning datasets suggest that with respect to types and subtypes of entity classes, the distinction between generic versus specific examples, and the consideration of events, ACE 2005 and BBN would be appropriate datasets for this project. In order to select one of them, I compared the number of instances per category as shown in Table 58. This is a relevant criterion because learning requires a substantial amount of examples per category. Note that in ACE, pronouns are also marked up as entities, and comprise about 14% of all annotated entities. This is very useful for reference resolution tasks, but for this project, I do not aim to classify pronouns as entities. Disregarding pronouns, BBN contains more than twelve times the number of entities that ACE offers. Therefore, I chose to use BBN as the learning dataset for this project.

Table 58: Quantitative comparison of suitable learning datasets

Category	**ACE 2005**	**Number of Examples**	**BBN**	**Number of Examples**
Agent	name	1,123	name	13,750
	nominal	2,111	descriptor	26,352
	pronoun	1,143		
	Subtotal (no pronoun)	3,234	Subtotal	40,102
Organization	name	887	name	19,450
	nominal	729	descriptor	30,244
	pronoun	182		
	Subtotal (no pronoun)	1,616	Subtotal	49,694
Location	name	127	name	1,088
	nominal	182		
	pronoun	24		
	Subtotal (no pronoun)	309	Subtotal	1,088
Facility	name	56	name	445
	nominal	343	nominal	2,570
	pronoun	45		
	Subtotal (no pronoun)	399	Subtotal	3,015
GPE	name	2,622	name	13,571
	nominal	527	nominal	1,835
	pronoun	382		
	Subtotal (no pronoun)	3,149	Subtotal	15,406
Vehicle	name	28	name	382
	nominal	183	nominal	1,223

	pronoun	27		
	Subtotal (no pronoun)	211	Subtotal	1,605
Weapon	name	15	name	21
	nominal	262	nominal	132
	pronoun	27		
	Subtotal (no pronoun)	277	Subtotal	153
Time		1,235		1,069
Money		94		11,097
Percent		17		5,976
Contact Info		2		40
Events	7 subtypes	1,557	3 subtypes	371
			Game	90
	Subtotal	1,557	Subtotal	461
Distinct classes	Values (3 subtypes)	165	Other named entities	9,448
			Other numerical entities	12,047
			Other temporal entities	20,676
Total	With Pronouns	14,094		
	Without Pronouns	**12,318**		**171,877**

Next, the categories in BBN had to be mapped to the meta-network categories. Table 59 shows the outcome of this process, which I performed in the following fashion: I picked one best match per category from the meta-network model by reviewing the descriptions in the BBN documentation, screening examples in BBN (last column in Table 59) and existing CASOS thesauri, and making sure that no category has too few examples (second column from the right in Table 59). The only category that I did not map onto a meta-network equivalent is "contact info: address", because a) this category has no good match in the meta-network and b) there are only four examples; two of which are overlapping with the class of "location: street".

Table 59: Category mapping from training data to category models

BBN	Mapping of BBN to Meta-Network				Example from BBN
Category name	**Category name**	**Subtype I**	**Subtype II**	**Examples /group**	
per_desc	agent	generic	na	26,352	activist
person	agent	specific	na	13,750	Arafat
org_desc:corporation	organization	generic	corporate	15,186	advertisers
org_desc:educational	organization	generic	educational	238	high school
org_desc:government	organization	generic	governmental	2,502	administration
org_desc:hospital	organization	generic	other		clinic
org_desc:hotel	organization	generic	other		hotel-casino
org_desc:museum	organization	generic	other		institution
org_desc:other	organization	generic	other	1,322	bar
org_desc:political	organization	generic	political	151	campaign
org_desc:religious	organization	generic	religious	51	church
organization:corporation	organization	specific	corporate	23,439	Occidental Petroleum Corp.

organization:educational	organization	specific	educational	366	Carnegie Mellon University
organization:government	organization	specific	governmental	4,629	Bank of Japan
organization:hospital	organization	specific	other		Harlem Hospital Center
organization:hotel	organization	specific	other		Ritz
organization:museum	organization	specific	other		Smithsonian Institute
organization:other	organization	specific	other	1,353	American Bar Association
organization:political	organization	specific	political	413	African National Congress
organization:religious	organization	specific	religious	44	Church of Scientology
norp:religion	org-att	specific	religious	88	Jewish
norp:nationality	org-att	specific	nationality	3,238	African
norp:other	org-att	specific	other	91	African-Americans
norp:political	org-att	specific	political	677	Communist
fac:airport	location	specific	facility		Heathrow
fac:attraction	location	specific	facility		Angel Fire
fac:bridge	location	specific	facility		Bay Bridge
fac:building	location	specific	facility		Andre Emmerich Gallery
fac:highway_street	location	specific	facility		101
fac:other	location	specific	facility	445	Auschwitz
fac_desc:airport	location	generic	facility		airport
fac_desc:attraction	location	generic	facility		aquarium
fac_desc:bridge	location	generic	facility		bridges
fac_desc:building	location	generic	facility		apartments
fac_desc:highway_street	location	generic	facility		circle
fac_desc:other	location	generic	facility	2,570	courtyard
gpe:city	location	specific	city	5,606	New York City
gpe:country	location	specific	country	5,079	Angola
gpe:other	location	specific	other		Bronx
gpe:state_province	location	specific	state-province	2,694	Alaska
gpe_desc:city	location	generic	city	377	capital
gpe_desc:country	location	generic	country	992	empire
gpe_desc:other	location	generic	other		borough
gpe_desc:state_province	location	generic	state-province	397	Baden-Wuerttemberg
location:border	location	specific	other		Four Corners
location:continent	location	specific	other		Africa
location:lake_sea_ocean	location	specific	other		Baltic Sea
location:other	location	specific	other		Alps
location:region	location	specific	other		Allegheny Mountains
location:river	location	specific	other	1,349	Amazon
animal	resource	na	animal	396	black widow
disease	resource	na	disease	317	cardiac condition
plant	resource	na	plant	194	cotton
product:other	resource	specific	product		Budweiser
product:vehicle	resource	specific	product		400 series
product:weapon	resource	specific	product	923	AH-64 Apache
product_desc:other	resource	generic	product		lifeboat
product_desc:vehicle	resource	generic	product		ambulance
product_desc:weapon	resource	generic	product	1,381	machine guns
substance:chemical	resource	na	substance		acid
substance:drug	resource	na	substance		cocaine

substance:food	resource	na	substance		bourbon
substance:nuclear	resource	na	substance		plutonium
substance:other	resource	na	substance	2,714	antibody
money	resource	na	money	11,097	$17
language	knowledge	specific	language	84	Arabic
law	knowledge	specific	law	382	425 U.S. 308
work_of_art:book	knowledge	specific	art		1984
work_of_art:other	knowledge	specific	art		60 Minutes
work_of_art:painting	knowledge	specific	art		Cemetery in the Snow
work_of_art:play	knowledge	specific	art		Death of a Salesman
work_of_art:song	knowledge	specific	art	721	I Can See Clearly Now
event:hurricane	event	specific	na		Hugo
event:other	event	specific	na		Big One
event:war	event	specific	na	371	French revolution
game	task	na	game	90	basketball
date:date	time	na	na		31-Mar-94
date:duration	time	na	na		10-month-long
date:other	time	na	na		annual
time	time	na	na	21,125	1 p.m. EST
cardinal	attribute	na	numerical		1.97
ordinal	attribute	na	numerical		200th
percent	attribute	na	numerical		0.30%
quantity:1d	attribute	na	numerical		1.2 miles
quantity:2d	attribute	na	numerical		8.2 by 11.7 inches
quantity:3d	attribute	na	numerical		1.6-liter
quantity:energy	attribute	na	numerical		900 megawatts
quantity:other	attribute	na	numerical		32-bit
quantity:speed	attribute	na	numerical		200 mph
quantity:temperature	attribute	na	numerical		321 degrees Fahrenheit
contact_info: other	attribute	na	numerical		ENG 23
Contact_info: phone	attribute	na	numerical		900-TELELAW
quantity:weight	attribute	na	numerical	18,059	2.5-ton
date:age	attribute	na	age	620	33

The BBN dataset had a few XML consistency issues that I fixed: four categories were defined in the BBN specification for which there were no examples in the annotated data. Eleven categories were not defined for BBN, but occurred in the annotated data with a total of 19 examples. I went through each of the examples and changed the category to what it should be according to the BBN documentation and the actual examples. One entity started as one type and ended as a different type; I adjusted that. Another issue with the data resulted from the fact that in XML data in general, a forward slash within an entity closes an XML tag prematurely. To avoid this issue, BBN places a forward slash right after a backward slash where applicable. This happens mainly for cardinal numbers, such as “1\/4 to 1\/2”, and organization, such as “Capital Cities\/ABC Inc.” However, a backward slash followed by forward slash is highly unlikely to be observed in new data. Therefore, I converted this structure into just a forward slash after parsing the XML files and prior to passing the input data to the learner.

3.3.2 Learning Technology and Selection of Feature Types

As a starting point for implementing the entity extractor, I used the CRF package as provided on the CRF project package (Sarawagi). This package offers a basic implementation of CRF, is highly adjustable, and allows for adding new features. The next challenge is to find a robust set of clues, also known as features, which bring together information about different characteristics of the data such that accuracy becomes high while predictions are robust. Robust here means that we need to avoid overfitting of the learned models to the idiosyncrasies of the learning data in order to ensure that the learner generalizes with high accuracy to new inference data. However, even though the feature set that will be chosen at the end of the feature selection process needs to support robustness, individuals features can be weak (Sarawagi, 2008).

Prior work has shown that in general, the following types of features are useful for entity extraction tasks: the identity of a token, i.e. the actual word or phrase, word surface features, orthographic features, syntax features, and external knowledge (Bikel et al., 1999; Borthwick et al., 1998; Cohen & Sarawagi, 2004; Florian et al., 2003; Mayfield, McNamee, & Piatko, 2003; McCallum & Li, 2003). In the following discussing of these features, I distinguish between “feature types” versus “features”, which are individual different clues per feature type.

3.3.2.1 Input Decomposition and Class Definition

Entity Extraction can be approached as a sequence labeling or a token labeling task. Token labeling means that for each individual word, two labels need to be predicted: 1) a boundary class label and 2) an entity class label or category. For example, for the entity “United Nations”, the predicted labels might be “begin, organization, specific” for “United”, and “end, organization, specific” for “Nations”. This task can be solved via one joint model for boundary and category, or two separate models for each label type. The advantage with the first approach is that there can be no conflicts between both label types. The disadvantage is that in the respective PGM, the number of classes, also known as states, and edges between states would be higher than with the second approach. As a result, fewer examples per class are available from the same training data. Furthermore, the higher complexity of the model leads to a higher time complexity for training. The advantages with the second approach are the higher number of examples per class, which also implies lower time requirements for learning. Furthermore, the features for boundary prediction and class label prediction can be tuned separately. The caveat is that both labels per token need to be combined in the end, which is highly likely to cause further loss in accuracy due to disagreements between both models.

With sequence labeling, one label gets predicted for each sequence, which can be a unigram or a multi-word expression. The same advantage and disadvantages as described above for the joint model of boundary and category prediction exist. Considering the outlined pros and cons, I chose to use the token labeling approach that predicts the boundary and category per token separately for the following reasons: the entity extractor built here is meant to support users in extracting two types of networks: one-mode networks, where all nodes are of the same type, and multi-mode networks, where nodes are instances of the meta-network categories. In order to extract nodes for one-mode networks, it is sufficient to correctly locate entities within their boundaries, but without assigning them to an entity class. Adding the detection of unigrams and bigrams as a stand-alone functionality to AutoMap would eliminate the need to identify these entities with alternative, computer supported techniques that require further manual vetting and selection (see section 5.2.2.1 for a description of how this is currently handled in AutoMap). This can be achieved with a prediction model that performs boundary detection only, which is the first reason for why I decided to construct a separate boundary prediction model. Next, in order to provide nodes for the construction of multi-modal networks, any located entities need further to be classified. This requires a second model for category prediction. In this process, however, nodes still need to be located as well. In order to keep the locating of nodes for one-mode networks and multi-mode networks in sync for the entity extraction method in general and for AutoMap in particular, I decided to use the same boundary prediction model for both situations, and to combine the boundary model with a class prediction model for building multi-mode networks (for details on combining both models see section 3.4.4).

Given the selected training data and the meta-network model, category labeling for this project can be based on four different category label models. These models are shown in Table 60. All of these models adhere to the meta-network ontology, but differ in the amount of granularity that they encoded in the entities (for details on the specific entity classes in each model see Table 59). Theoretically, entity class model 4, which is the most complex or detailed one as it specifies the meta-network category, specificity and subtype of each entity, can be reduced to each of the other entity class models. However, due to the model complexity and thus the lower number of training instances per category, the model might not perform as well as the simpler ones. This would mean a loss of accuracy or practical usefulness for the end-user. The same argument can be made for reducing entity class models 2 (category and specificity) and 3 (category and subtype) to entity class model 1 (category only). My assumption here is that higher complexity leads to lower accuracy. I am reporting on the outcome of testing this hypothesis in the results section. The choice for a specific model has another aspect to it: for practical purposes, different datasets and research questions might require different levels of detail such that we cannot anticipate which model would be most useful. Thus, each of the models could be suitable for text coding in AutoMap, and would expand the current scope of capabilities of this tool. Thus, we decided to generate all four options, and to report on their accuracy and robustness so that users can pick the model that best serves their needs; potentially trading off accuracy for granularity.

Table 60: Entity class model definition

	Category name (meta-network classes)	**Subtype I (generic vs. specific)**	**Subtype II (attributes per class)**	**Example**
Entity class model 1	x			agent
Entity class model 2	x	x		agent, specific
Entity class model 3	x		x	agent, political
Entity class model 4	x	x	x	agent, specific, political

Table 61 reports on the complexity of the token labeling approaches (separate versus joint models for boundary and category) and the class label models in terms of the number of classes and edges and run time. These tests were performed by learning with 80% of the data (4 holdout folds) and making predictions on the remaining 20% of the data (1 holdout fold) for two different, but not all five holdout folds, and averaging the results. A more complete description of the evaluation routine is provided in section 3.4.1. Each of the tested holdout folds has about 43,000 labeled tokens. The runtime was measured with the baseline feature set that is explained in section 3.3.2.2. The time needed for a single iteration of the CRF varies greatly depending on the model complexity[10]: for boundary detection, it is only one minute, while for joint prediction of boundary and category with entity class model 4, it is 175 minutes. As reported in section 3.4.4 in more detail, 300 iterations is a rate at which results start to stabilize. This rate would require over a month of runtime for the most complex model for the joint prediction option. However, during the feature testing and selection stage, it is crucial to test the contribution of each feature type to accuracy separately, to then modify or drop features accordingly, and to repeat this process as often as necessary. The token level approach, especially one that breaks boundary and category prediction into separate tasks, supports this need better than the alternative approach. This fact is the second reason for why I chose the token level approach that involves a model for boundary and category prediction each. However, I present extraction results for both token labeling approaches with a low iteration rate in order to clarify on the difference in accuracy.

[10] All experiments described in this chapter were run on a total of three different machines with 64 bit operating systems. One machine had 256 GB of RAM and 24 processors, the other two machines had 512 GB of RAM and 64 processors.

Table 61: Token labeling approaches: complexity per model*

Token labeling approach	Size and Runtime costs	Boundary Model	Entity class model 1	Entity class model 2	Entity class model 3	Entity class model 4
Separate models for boundary and class	Number of States Number of Edges Runtime: Min. per iteration Runtime for 300 iterations	5 25 1 5 hours	11 121 3.5 17.5	16 256 6 1.25 days	32 1,024 15 3.1 days	45 2,025 24 5 days
Joint model for boundary and class	Number of States Number of Edges Runtime: Min. per iteration Runtime for 300 iterations	n.a.	41 1,681 17 3.5 days	60 3,600 31 6.5 days	121 14,641 126 26.3 days	155 24,025 175 36.5 days

*holdout folds 1,3, number of states and edges for sequence level from holdout fold 3

3.3.2.2 Baseline Features

The CRF project package contains various feature types. The following eight features are the ones that I considered as being potentially relevant for establishing a baseline for this project:

1. *Word Features:* Identity per token.
2. *Word Score Features:* The log of the number of tokens with a certain label over the number of all tokens with that label.
3. *Edge Features*: Information about transitions between states.
4. *Start Features:* Active when current state is a start state.
5. *End Features:* Activate when current state is an end state.
6. *Unknown Feature:* Active for token not observed during training.
7. *Known In Other State Feature:* Active when a token was not observed in a particular state, but in other states with more than a minimum threshold frequency.
8. *Regex Features*: A collection of multiple orthographic characteristics and regular expressions per token.

All of these features are implemented on a per state basis, except for the first feature, which is implemented on a per token level. Overall, these features represent common features for information extraction tasks that are solved via machine learning methods, especially those that use PGM with Markov properties (Bikel et al., 1999; Diesner & Carley, 2008b; Ratinov & Roth, 2009). This particularly applies to the edge features, the start and end features, and the unknown feature.

3.3.2.3 Syntax Features

In order to identify the part of speech (POS) for each token, I use the POS tagger that I had previously built for AutoMap (Diesner & Carley, 2008b). This tagger implements a HMM via the Vitberi algorithm, operates on the sentence level, and tags every sequence of characters that is composed of any combination of letters, numbers, dashes, ampersands, dollar symbols, and single hyphens. The latter mainly serves as genitive markers. Any token that does not match this pattern is disregarded for tagging, including hyphens composed of two single hyphens. The tagger achieves an accuracy of over 93% on predicting two different tag sets: the Penn Treebank (PTB) tag set with 36 tags, and a set where the PTB tags are aggregated into more general tags, such as all verb forms to "verb" (for the mapping from PTB to the aggregated tag set see the Appendix in Diesner & Carley, 2008b). I refer to these tag sets as "full" and "aggregated", respectively, in the following.

Using the tagger for this project revealed two issues: First, the tagger predicts two categories that do occur in the training data that the tagger was built based upon, i.e. PTB 3 (P. M. Mitchell, Santorini, & Marcinkiewicz, 1993), but that are not defined for the full PTB tagset. Specifically, the tag "JJSS" should rather be "JJS", and "PRP$R" should be "PRP$P". This problem was noted by others before (Pereira, 2004), but was not spotted when building the AutoMap POS tagger. In order to find out if this glitch matters, I mapped the two undefined categories onto the ones they truly should be and tested the impact on the entity extraction accuracy. The results as shown in Table 62 suggest that this ex post factum fix hurts prediction accuracy, mainly by lowering recall. This is because in the POS training data, the undefined tags were assigned to one different term each, such that the resulting tagger would put these words into separate classes of their own. In order to keep the entity extractor in sync with AutoMap, which uses the POS tagger that contains the additional two categories, I decided to not to keep this change for further work on this project. Ultimately, this issue can be solved by retraining the tagger.

Table 62: Impact of Part of Speech tag fix on accuracy*

	Boundary Prediction		**Class Prediction**	
	original	fixed	original	fixed
Precision	88.1%	88.4%	85.7%	85.7%
Recall	85.7%	85.1%	81.2%	81.0%
F	**86.9%**	**86.7%**	**83.4%**	**83.3%**

*Iteration Rate 200, Class model 1

Second, when I screened the results of POS tagging of the tokens in BBN, I realized that most tagging errors applied to numbers, especially percentages, which were wrongfully assigned to classes other than the numbers class. However, in the BBN data, most of the tokens that involve digits truly are numbers. Thus, I made another ex post factum change to the POS tagger: any token that contains a digit is tagged as a number, i.e. as "CD" for the PTB full set, and as "NUM" for the aggregated set. I kept this change for learning.

Part of speech can be used as a feature for CRF as a) a per state feature, or b) a per state and per word feature. Which of these two options and which of the two available POS tag sets achieve higher accuracy rates is shown in the results section.

3.3.2.4 Lexical Features

Prior research has shown that the accuracy of entity extraction can be increased by adding features that use external knowledge sources such as a lookup dictionary (Brown, Desouza, Mercer, Pietra, & Lai, 1992; Bunescu et al., 2005; Cohen & Sarawagi, 2004; Ratinov & Roth, 2009). In fact, several of the potential trainings sets discussed in this chapter include gazetteer data as additional files. Using dictionaries has also been shown to help with domain adaption, i.e. adapting an extractor from the training data domain to other domains for conducting inference (Ciaramita & Altun, 2005).

For this project, I use the thesaurus that I prepared as described in detail in section 5.2.2.1.1 as a dictionary. This thesaurus contains 169,791 entries and is herein referred to as the "master thesaurus". The left hand side of the thesaurus contains potential text level entries, and the right hand side has the related meta-network category. Of those entries, 59.6% are locations. However, this category includes plenty of noisy entries, which mainly result from scraping the web without careful cleaning the retrieved hits, and adding stemmed versions and foreign translations of location to the thesaurus; some of which might be valid English words that would rather belong into different meta-network categories. Both of these routines were performed by others before I took over work on the master thesaurus. I fixed many of those issues as described in section 5.2.2.1.1. However, I neither removed the translations nor locations that were unknown to me, but sounded like valid entries. Since runtime costs increase with the size of the thesaurus, but many of these location entries are unlikely to occur in new text data, I built a reduced version of the master thesaurus as follows: I took out all locations (169,791 entries) and replaced them with just the names of all countries and capitals in the world (439 entries) as provided in (Bureau_of_Intelligence_and_Research, 2011). The resulting thesaurus contains a total of 69,067 entries and is 59.3 % smaller than the original master thesaurus. I refer to this thesaurus as the "reduced master thesaurus".

Building upon prior work and extending it with new lexical features, I added the following lexical features to the CRF implementation:

1. *Is in Dictionary Feature:* Activated if token matches complete content of left hand side entry in thesaurus. Executed on the unigram level. Implemented per state. This feature is motivated by (Ciaramita & Altun, 2005).

2. *Is in Dictionary per Word Feature:* Same as above, but implemented per state and per word.
3. *Occurs in Dictionary Feature:* More relaxed version of the “Is in Dictionary Feature”. Activated if token matches any part of the content of left hand side entry in thesaurus. Matches on token level among unigrams and within N-grams are valid. Implemented per state. This feature is motivated by Cohen and Sarawagi (2004).
4. *Position in Dictionary Feature:* If token occurs in dictionary, this feature records the position of a token in the left hand side entry of the thesaurus. Matches among unigrams and within n-grams are valid. Positions available are begin, inside, end, and unique. Example: if the token is “House” and the thesaurus contains “White House”, then “House = end” gets recorded. Implemented per state. This feature is motivated by Cohen and Sarawagi (2004).
5. *Position in Dictionary per Word Feature:* Same as above, but implemented per state and per word.
6. *Category Feature:* If token occurs in left hand side entry of thesaurus, this feature records the meta-network category of that token. Matches among unigrams and within n-grams are valid. Implemented per state.
7. *Category per Word Feature:* Same as above, but implemented per state and word.

Cohen and Sarawagi (2004) have shown that using soft matches instead of exact matches of tokens to dictionary entries further increases accuracy. However, the thesauri I use already contain grammatical and lexical variations of words, including inflexions, conjugations, morphemes, abbreviations, and synonyms. Further computing string similarities between text tokens and the dictionary entries might enable the consideration of more token variants than those already provided in the thesauri, but might also pick up on false positives. Moreover, computing string distance metrics adds significant time costs to the learning process, especially for dictionaries as large as the ones used here. For the given reasons, I only consider hard matches between text tokens and dictionary entries, but compute a variety of dictionary features that aim to capture different characteristics of the thesaurus entries.

3.3.3 Experimental Design

Table 63 gives an overview on the feature types or variables that need to be tested for their individual and combined contribution to extraction quality. This table also specifies the variables’ value ranges that I consider potentially useful for this project. Testing all combinations of the values of the selected feature types would result in an 8*9*2*5*2*2*2*7 = 40,320 design. Doing these experiments would be an overkill for this project because not all combinations are meaningful, and many of them can be ruled out once the best value for a specific variable has been identified. Thus, I mainly conduct experiments to identify the best value per parameter, and then incrementally combine them across parameters.

Table 63: Experimental design: variables and values

Variable	Values								
Baseline	Word Features	Word Score Feature	Edge Features	Start Features	End Features	Un-known Feature	Known in other state Fea.	Regex Features	
Iteration Rate	100	200	300	400	500	600	700	800	900
Token Labeling	Separate models for boundary and class	Joint model for boundary and class							
Class label model	Boundary Model	Entity class model 1	Entity class model 2	Entity class model 3	Entity class model 4				
Syntax Features	PTB full	PTB aggregated							
	POS per state	POS per word							
Lexical Features	Full master thesaurus	Reduced master thesaurus							
	Is in Dictionary Feature	Is in Dictionary per Word Feature	Occurs in Dictionary Feature	Position in Dictionary Feature	Position in Dictionary per Word Feature	Category Feature	Category per Word Feature		

3.4 Results

3.4.1 Evaluation Method and Metrics

The accuracy rates presented in this section were obtained by performing k-fold cross validations: I split up the BBN data into five chunks, also known as folds, of about equal size. The folds are static, i.e. the same files stay in the same bucket for all experiments. For each run, all folds expect for the holdout folds are used for training a prediction model. During evaluation, the learned model is applied to the holdout fold, and each deviation from the original tag per token in the holdout fold (ground truth) is recorded as an error. At the end of all runs per experiments, where the number of runs equals k, the obtained accuracy rates are averaged. No fold is ever used for training and evaluation in the same run. Ideally, one would iterate through each of the five folds as being the holdout fold once per experimental condition (5-k cross validation). This strategy is used for assessing the accuracy of the final models. Practically, the experiments were constrained by the computing resources that were available to me and the time costs for experiments. Therefore, I use a reduced approach for assessing the accuracy rates for the values per variable: I perform two runs per experimental condition with two randomly selected holdout sets, which were folds 1 and 3.

3.4.2 Points of Comparison for Accuracy Rates

To the best of our knowledge, no other group has used BBN to predict the meta-matrix categories specifically. Therefore, I have no precise external point of comparison for the accuracy rates that will be obtained. However, results from the main Named Entity Extraction initiatives are applicable points of comparison: in ConLL 2003, the Named Entity task involved extracting the boundary and category labels for the classes of person, organization and location. The top five systems achieved F-measures of 85% and more; with the best system having an F value of 88.7% (CoNLL-2003, 2003; Florian et al., 2003). In MUC7, the categories to predict were more similar to BBN that those used in CoNLL 2003, and in fact, BBN data was part of this task (for details see Table 54 and Table 55). The top two systems in MUC7 achieved F-values of 91.6% and 94.4%, and four more systems had F-values of more than 85% (MUC7, 2001). The goal with this project is not to beat these benchmark values, but to stay in the range of state of the art performance values by using cutting edge methods and technologies, and also leveraging on routines (e.g. POS tagging) and material (e.g. lookup dictionary) that I have developed for AutoMap and CASOS. These routines and materials are an integral part of current tools and research projects that we have developed and conducted, respectively.

Previously, we have applied CRF to BBN to train a model that predicts a class label per token with an accuracy rate of 82.7% (Diesner & Carley, 2008a). This model differs from the ones build in this project in the following ways: First, it only operates on the unigram level, i.e. multi-word expressions are not retrieved as such. In other words, no boundary detection is performed. Second, it uses entity class model 1, i.e. meta-network categories only without further attributes. Third, it considers a smaller number of the categories available in BBN (details on the mapping of BBN categories to meta-network categories are provided in Table 1 in (Diesner & Carley, 2008a). The goal with this project is to improve on this baseline in multiple ways: first, to extract unigrams as well as N-grams, second, to extract entities that adhere to more complex entity class models; third, to capture attributes per entities; and forth, to improve accuracy.

3.4.3 Baseline

As the results in Table 64 show, six of the eight baseline feature types contribute to accuracy. The “known in other state” feature has no impact. The “word score” feature reduces accuracy by a few percentage points. The ranking of how much the feature types impact accuracy is the same for the three most useful feature types for both, boundary and category prediction. The “word identity” feature is by far the strongest clue. Information about transitions is also greatly helpful. From this point on, the features that are not contributing to accuracy are excluded from the feature set such that the baseline consists of six feature types.

Table 64: Accuracy loss due to elimination of each single baseline feature*

Boundary	All Baseline Features	Word	Edge	Regex	Start	End	Unknown	Other State	Word Score
Precision	84.5%	-28.3%	-19.9%	-2.9%	-0.3%	0.0%	-0.1%	0.0%	3.9%
Recall	83.7%	-38.5%	-24.5%	-6.0%	-1.6%	-2.0%	-2.7%	0.0%	3.2%
F	84.1%	-34.0%	-22.3%	-4.5%	-1.0%	-1.0%	-1.4%	0.0%	3.5%
Rank (based on F, 1=best)		1	2	3	5	6	4	no contributor	no contributor
Class	**All Baseline Features**	**Word**	**Edge**	**Regex**	**Start**	**End**	**Unknown**	**Other State**	**Word Score**
Precision	84.8%	-31.1%	-10.5%	-3.6%	-0.1%	-1.7%	-1.0%	0.0%	2.6%
Recall	82.3%	-46.9%	-11.9%	-2.3%	-0.7%	-2.2%	0.1%	0.0%	1.9%
F	83.5%	-41.3%	-11.3%	-2.9%	-0.4%	-2.0%	-0.4%	0.0%	2.2%
Rank (based on F, 1=best)		1	2	3	5	4	6	no contributor	no contributor

*Iteration rate = 300, class model 2, holdout folds: 1,3, Class

3.4.4 Iteration Rate and Input Decomposition

Increasing the number of iterations leads to substantial gains in accuracy up to an iteration rate of about 500, where gains start to become minimal, as shown in Table 64. In Table 64, the last horizontal row in each section shoes the change rate in F as the iteration rate is increased by 100. Accuracy starts to drop from about 700 iterations on. Precision is higher than recall and benefits less form increasing the iteration rate than recall does, though this effect decrease as the iteration rate is increased.

Figure 10 illustrates this effect for a particular holdout set: the number of tokens retrieved and tokens correctly classified increases approximately by the same amount per iteration rate. For practical purposes, however, recall is more important than precision as retrieved yet misclassified entities (false positives) might be suitable fits for alternative categories. Overall, the results support the strategy of using an iteration rate of 300 for further testing of the impact of features since the results are fairly robust at this point.

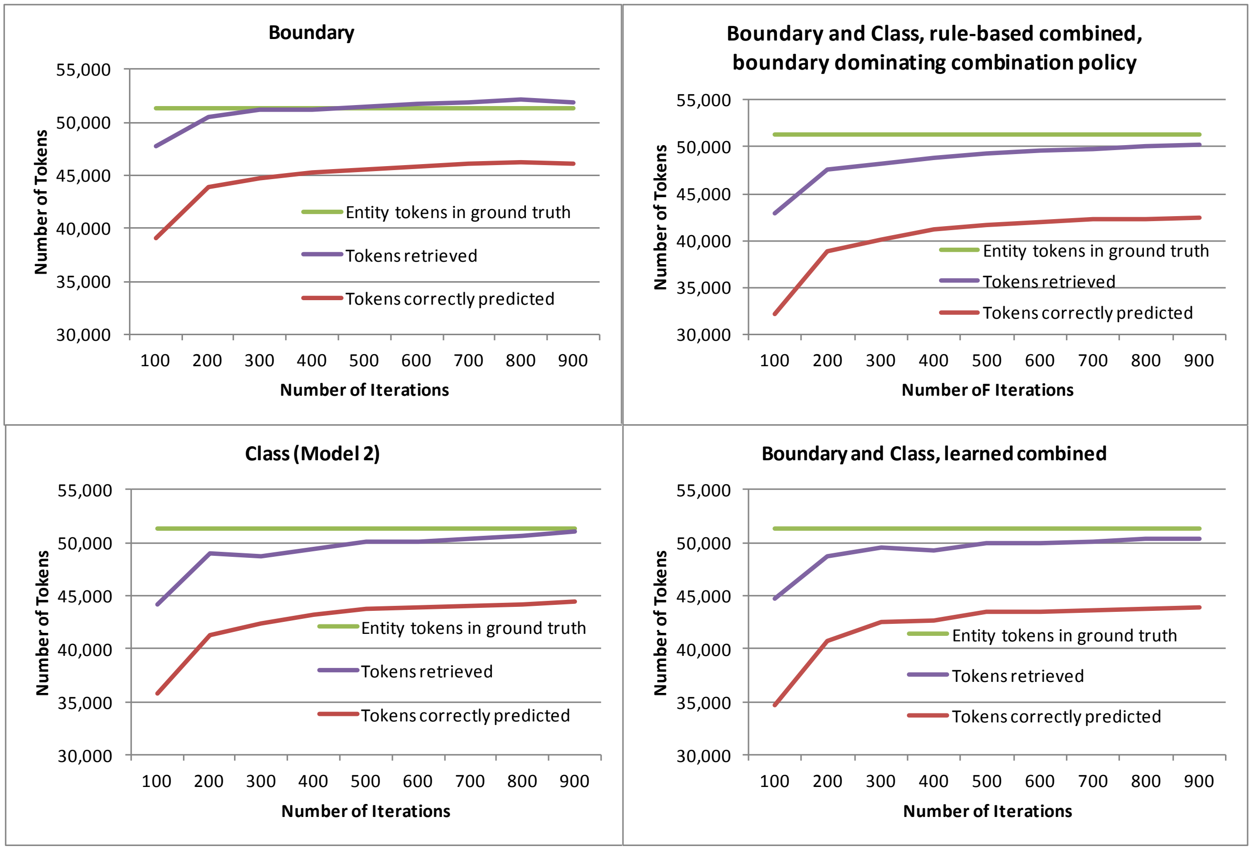
Table 65: Impact of iteration rate on accuracy*

	Iteration Rate								
Boundary	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**
Precision	82.8%	87.3%	88.4%	89.0%	89.1%	89.3%	89.4%	89.6%	89.5%
Recall	77.6%	85.3%	86.9%	88.1%	88.9%	89.3%	89.6%	89.6%	89.9%
F	80.1%	86.3%	87.6%	88.5%	89.0%	89.3%	89.5%	89.6%	89.7%
Change Rate in F		6.2%	1.3%	0.9%	0.5%	0.3%	0.3%	0.0%	0.1%
Class (Model 2)									
Precision	82.4%	86.0%	87.9%	88.4%	88.4%	88.6%	88.5%	88.4%	88.2%
Recall	70.0%	80.6%	82.9%	84.3%	85.1%	85.6%	86.1%	86.3%	86.6%
F	75.7%	83.2%	85.3%	86.3%	86.7%	87.1%	87.2%	87.3%	87.4%
Change Rate in F		7.5%	2.2%	0.9%	0.4%	0.4%	0.1%	0.1%	0.1%
Boundary & Class	Rule-based combination of separately learned models, boundary dominates class								
Precision	76.4%	82.7%	84.4%	85.1%	85.3%	85.6%	85.5%	85.5%	85.3%
Recall	63.6%	75.8%	78.5%	80.2%	81.3%	81.9%	82.4%	82.3%	82.8%
F	69.4%	79.1%	81.3%	82.6%	83.2%	83.7%	83.9%	83.9%	84.0%
Change Rate in F		9.7%	2.3%	1.3%	0.6%	0.5%	0.2%	0.0%	0.2%
Boundary & Class	Rule-based combination of separately learned models, class dominates boundary								
Precision	75.3%	79.3%	82.0%	82.7%	82.7%	83.0%	83.0%	83.0%	82.7%
Recall	64.0%	74.3%	77.4%	78.9%	79.6%	80.2%	80.7%	81.0%	81.2%
F	69.2%	76.7%	79.6%	80.8%	81.2%	81.6%	81.8%	82.0%	81.9%
Change Rate in F		7.5%	2.9%	1.1%	0.4%	0.5%	0.2%	0.1%	0.0%
Boundary & Class	Learned joint model								
Precision	78.3%	84.5%	86.7%	87.8%	88.1%	88.2%	88.0%	88.1%	88.2%
Recall	67.1%	79.2%	82.6%	83.4%	84.9%	84.9%	85.5%	85.7%	85.9%
F	72.3%	81.8%	84.6%	85.6%	86.5%	86.5%	86.7%	86.9%	87.0%
Change Rate in F		9.5%	2.8%	1.0%	0.9%	0.0%	0.2%	0.2%	-0.6%

* Holdout folds 1,3

With respect to the results for input decomposition, the results in Table 65 suggest that when separate models are learned for boundary and category prediction, boundary prediction is over 2% more accurate than category prediction. This seems intuitive since the boundary model contains less than half the number of labels of the entity class model (in this case Nr. 2) does. Learning a joint model for boundary and category prediction (last horizontal section in Table 65) is slightly less accurate than learning separate models for both types of prediction prior to consolidating them. This difference becomes smaller as the iteration rate increases; at 500 iterations it is 2.5% and 0.2% in comparison to boundary prediction and class prediction, respectively. However, when separate models are learned for boundary and category prediction, these models need to be merged in the end, and accuracy assessment needs to be performed again on the joint models. My results show that either approach of merging as explained right below leads to accuracy rates that are about 3% and more less accurate than those obtained with the joint model. However, I argue that learning boundaries and category labels with separate models leads to more robust final models because there is much more training data available for each class. Also, learning the joint model took four times as long (10.8 days at 500 iterations) than the separate models did (2.1 days). Since we aim for high generalizability of the models, I chose to stick with this more robust solution.

Figure 10: Diminishing returns: Impact of iteration rate on accuracy*



* Class model 2, holdout fold 1

The decision to work with separately learned models for boundary and category prediction implies that once both types of models have been generated, they need to be combined before inference can happen. This combination needs to be done such that we obtain a) both, a boundary label and a class label, for each token and b) consistent labels, especially for multi-word units. Table 66 provides an overview on the discrepancies that that can occur.

I developed and implemented a rule based approach for combining these models and resolving any discrepancies between them by considering all logically possible mismatches and suggesting a solution for each of them, and using a data driven approach for checking the learned baseline models for the characteristics of these discrepancies and testing the impact of any suggested solution. The outcome of this process, i.e. the resulting rule set, is shown in Table 66. The developed rule set is based on two different policies for handling mismatches between boundary and class labels: 1) boundary prediction dominates class prediction, and 2) class prediction dominates boundary prediction:

Boundary prediction dominates category prediction: If there is a class label but no boundary label with the value of begin, inside or end, the token is not considered as an entity. If the class labels in a multi-word unit according to boundary prediction are not coherent, I assign the most frequent label (other than none) to all tokens in that expression. In the case of a tie, the first category is picked. For cases in which boundary prediction finds a unigram but no class label is suggested, I tested two strategies: not considering the token as a relevant entity al together, or assigning the token to the most frequent class label. My error analysis of the outcome suggested that the errors fall with almost equal frequency into three categories: 1) being a token of the type of the most frequent type of entity class, 2) being a token of some other type of entity class, or 3) being a false positive according to boundary prediction. Case 2 occurred slightly more frequently than case one. Therefore, I chose to assign no class label to unigrams that lack a class label and converting these entities to the "outside" boundary condition.

Category prediction dominates boundary prediction: If a token has a class label other than none, but the token right before and after do not, and the boundary label for this token is outside or part of a multi-word unit, the boundary label is set to "unigram". If the sequencing of boundary labels does not coincide with a multi-word unit according to class label prediction, the boundary labels are adjusted accordingly. Note that with this policy, mismatching unigrams are preserved, while with the first policy, they are lost, which gives the second policy a potential advantage over the first one.

Testing both policies empirically suggests that letting the using the policy where the boundary label dominates the category label returns slightly more accurate results (1% and less). This finding seems intuitive because boundary prediction is overall more accurate than class label prediction. Cases in which the category dominating policy preserved unigrams led to significant ratios of truly false hits, which diminished the potential gains from this strategy.

The rule-based procedure described in this section was only used for accuracy assessment throughout the results section of this chapter. For integrating the entity extractor into an end-user software product, a more permissive approach was chosen in order to allow for higher recall. This approach is explained in section 4.

Table 66: Rules for model combination depending on combination policy

Policy	**Case**	**Learned Labels**		**Combination Result**	
Boundary dominates Category		Boundary	Class	Boundary	Class
	1	none	positive token (i.e. category not none)	none	none
	2	unigram	none	none	none
	3	N-gram	all tokens none	none	none
		N-gram	different category labels, at least one positive token	N-gram as learned	majority class label other than none, ties broken alphabetically
Category dominates Boundary	4	unigram	none	none	none
	5	none, begin, inside, end	positive token	unigram	positive token as learned
	6	inconsistent with class label sequence, incl. one to all boundary labels equal none	positive N-gram	proper N-gram	positive N-gram as learned

These and many other results for the impact of individual feature type values on accuracy were obtained by averaging the outcomes of cross-validations with holdout sets 1 and 3. In order to verify that these two folds are not outliers, which would impact the drawn conclusions and subsequent modeling decisions, I present a snapshot of sample sizes, number of features, and accuracy rates for all holdout sets for a constant iteration rate in Table 68. These numbers show that basically all five folds are similar in size, and lead to similar accuracy rates; with a variation in F of about 0.4% for boundary prediction and 1.6% for class prediction. Also note that the number of features is between 50,000 and 51,250 for class prediction, and between 53,500 and 54,500 for boundary prediction. This means that with only six baseline feature types, a large number of features is generated; with most of them being word features. This also means that for boundary prediction, which involves 5 states and 25 edges, more features are generated than for class prediction, which has 16 states and 256 edges for this entity class model. The reason for this counterintuitive effect is that with fewer classes, the learning data is less sparse such that more useful features might be found.

Table 67: Size and accuracy per holdout set at constant iteration rate

Measures	**Holdout Set: 1**	**2**	**3**	**4**	**5**
	Boundary				
Number of Entity Tokens	43380	43467	42937	43078	43652
Number of Features	54122	54204	53607	53737	54455
Precision	86.9%	87.3%	87.7%	87.8%	87.4%
Recall	85.4%	85.6%	85.2%	85.4%	85.3%
F	86.2%	86.4%	86.4%	86.6%	86.3%
	Class (Model 2)				
Number of Entity Tokens	43380	43467	42937	43078	43652
Number of Features	50824	50944	50355	50476	51252
Precision	84.4%	86.7%	87.6%	87.6%	86.8%
Recall	80.5%	79.9%	80.7%	80.2%	80.2%
F	82.4%	83.1%	84.0%	83.7%	83.4%

*Iteration rate = 200, holdout folds: 1,3

3.4.5 Syntax Features and Entity Class Models

In general, most features can be implemented on a a) per state or b) per word and state basis. Table 68 shows a comparison of these two options for the part of speech tags feature type. The per state approach leads to a slightly higher accuracy (less than 1%) with less than half the number of features generated, i.e. the per state option is more efficient and more robust. Therefore, this option is used for further work.

Table 68: Impact of Part of Speech tag feature implementation approach on accuracy*

POS Feature		Boundary		Class	
Implementation	Iteration Rate	200	400	200	400
Per State	Precision	88.1%	89.3%	85.7%	88.4%
	Recall	85.7%	88.4%	82.1%	84.8%
	F	**86.9%**	**88.9%**	**83.8%**	**86.6%**
Per Word and State	Precision	87.7%	88.8%	86.5%	88.4%
	Recall	85.1%	88.1%	80.0%	84.5%
	F	**86.4%**	**88.5%**	**83.1%**	**86.4%**

* holdout folds: 1,3, Class model 2

The results for the impact of using part of speech as a feature type (Table 69) suggest that both, the aggregated as well as the full tag set, have a small positive impact on accuracy rates. The full tag set leads to higher gains in accuracy over the baseline than the aggregated set does for boundary detection and all entity class models except for model 4, where the results for both tag set tie.

Table 69: Impact of Part of Speech tag features and entity class models (models sorted by accuracy) on accuracy*

Assessment Metrics	BL	POS Agg	POS Full
	Boundary		
Precision	88.4%	89.1%	89.1%
Recall	86.9%	86.5%	87.5%
F	87.6%	87.8%	88.3%
Change in F from Baseline (BL) to POS		**0.2%**	**0.7%**
	Entity class model 2 (meta network category + gen/spec)		
Precision	87.9%	86.9%	87.0%
Recall	82.9%	83.7%	84.3%
F	85.3%	85.3%	85.6%
Change in F from BL to POS		**-0.1%**	**0.2%**
Diff. in F over next less accurate class model	**1.3%**		**0.6%**
	Entity class model 1 (meta network category)		
Precision	85.5%	86.5%	86.5%
Recall	82.6%	82.8%	83.5%
F	84.0%	84.6%	85.0%
Change in F from BL to POS		**0.6%**	**1.0%**
Diff. in F over next less accurate class model	**1.0%**		**1.4%**
	Entity class model 4 (meta nw. cat. + gen/spec + subtype)		
Precision	85.3%	85.5%	85.1%
Recall	80.9%	81.9%	82.1%
F	83.0%	83.6%	83.6%
Change in F from BL to POS		**0.6%**	**0.6%**
Diff. in F over next less accurate class model	**0.9%**		**0.5%**
	Entity class model 3 (meta network category)		
Precision	83.5%	84.4%	84.6%
Recall	80.9%	81.2%	81.5%
F	82.2%	82.8%	83.1%
Change in F from BL to POS		**0.6%**	**0.9%**

* Iteration rate = 300, holdout folds: 1,3

With respect to entity labeling according to the four different entity class models as defined in Table 60, the results in Table 69 indicate that accuracy rates do not necessarily drop as the complexity of the models, i.e. the number of states and edges, increases. In fact, the second smallest model (entity class model 2, category and specificity), performs best. Also, the most complex model (model 4, category, specificity, subtype) outperforms model 3 (category, subtype). Moreover, the accuracy differences between the entity class models are fairly small (2.5% for the widest gap after POS tagging), even though the model complexities are very different (the number of classes differ by a factor of about 4 between the largest and the smallest entity class model). Based on these results I reject my hypothesis that greater model complexity leads to lower accuracy rates.

3.4.6 Lexical Features

Adding lexical or dictionary features boost accuracy by up to 4% (Table 70). However, only four of the seven dictionary features defined and tested for this project have a robust, positive impact on accuracy across dictionaries (full versus reduced master thesaurus) and prediction models (boundary versus category). These are the "Is in Dictionary per Word Feature (by far the strongest feature), Category Feature, Category per Word Feature, and Position in Dictionary per Word Feature. The Position in Dictionary Feature returns the exact same results as the Is in Dictionary Feature. The same is true for the Position in Dictionary Feature per Word and the Category Feature. Therefore, both Position in Dictionary features are excluded from here on.

For most of the tested conditions, using the full master thesaurus as a dictionary leads to slightly better results than the using the reduced master thesaurus (0.4% on average for the selected dictionary features). However, the full master contains more than twice as many entries as the reduced one does, but hardly leads to more than twice as much accuracy gain. Therefore, I chose to use the reduced master thesaurus as well as the Is in Dictionary per Word Feature, Category Feature, and the Category per Word Feature for further work.

Table 70: Impact of dictionaries and dictionary features on accuracy

Features	Baseline	Is in Dictionary	Is in Dict. per Word	Category Feature	Category per Word	Occurs in Dictionary
	Boundary, Reduced Master Thesaurus					
Precision	88.4%	88.6%	92.1%	88.5%	88.5%	88.5%
Recall	86.9%	87.2%	90.6%	87.5%	87.3%	86.6%
F	87.6%	87.9%	91.3%	88.0%	87.9%	87.5%
Difference to BL**		0.31%	**3.71%**	**0.42%**	**0.32%**	-0.10%
	Boundary, Full Master Thesaurus					
Precision	88.4%	89.0%	92.1%	88.9%	88.6%	88.5%
Recall	86.9%	86.7%	91.1%	87.9%	87.7%	87.0%
F	87.6%	87.8%	91.6%	88.4%	88.2%	87.7%
Difference to BL**		0.22%	**3.98%**	**0.80%**	**0.56%**	0.12%
	Class (Model 2), Reduced Master Thesaurus					
Precision	87.9%	87.3%	91.1%	88.0%	87.8%	88.0%
Recall	82.9%	82.6%	86.3%	84.0%	83.4%	82.5%
F	85.3%	84.9%	88.6%	85.9%	85.5%	85.1%
Difference to BL**		-0.48%	**3.27%**	**0.56%**	**0.18%**	-0.21%
	Class (Model 2), Full Master Thesaurus					
Precision	87.9%	87.6%	91.4%	87.7%	87.8%	87.8%
Recall	82.9%	82.7%	87.3%	84.0%	84.1%	82.5%
F	85.3%	85.1%	89.3%	85.8%	85.9%	85.1%
Difference to BL**		-0.27%	**3.92%**	**0.49%**	**0.54%**	-0.28%

* Iteration rate = 300, holdout folds: 1,3

** Bold if gain over BL for both holdout folds

3.4.7 Final Feature Set

Based on the presented results from the tests of the impact of iteration rate, input decomposition, syntax features and lexical features, the feature set shown in Table 71 was used for constructing the model to be integrated into AutoMap.

Table 71: Final feature set for prediction models (active feature types in black, feature types not chosen in gray)

Variable	Values								
Baseline	Word Features	Word Score Feature	Edge Features	Start Features	End Features	Un-known Feature	Known in other state Fea.	Regex Features	
Iteration Rate	100	200	300	400	500	600	700	800	900
Decom-position	Token Level	Sequence Level							
Class label model	Boundary Model	Entity class model 1	Entity class model 2	Entity class model 3	Entity class model 4				
Syntax Features	PTB full	PTB aggregated							
	POS per state	POS per word							
Lexical Features	Full master thesaurus	Reduced master thesaurus							
	Is in Dictionary Feature	Is in Dictionary per Word Feature	Occurs in Dictionary Feature	Position in Dictionary Feature	Position in Dictionary per Word Feature	Category Feature	Category per Word Feature		

For these experiments, a 5-fold cross-validation was conducted. The results in Table 72 show the accuracy rates for the entity class models with the final feature type configuration. Overall, the performance of the combined boundary and class label models is very similar across the different class label models; with 1.4% difference at most. This indicates that large differences in model complexity have little impact on accuracy. The results also confirm the previously identified ranking of models based on accuracy, with the least complex model being outperformed by the next complex model, and the most complex model being more accurate than the next less complex one. Moreover, the obtained results (accuracy between 87.5% and 88.8% for the combined models) are comparable to alternative top performing systems, where accuracy rates typically range in the 80ies and lower 90ies (see for example Florian et al., 2003; MUC7, 2001). Furthermore, the achieved rates are 6% to 7% higher than the ones achieved with the previous entity extractor in AutoMap, which used a less complex category model (Diesner & Carley, 2008a).

Table 72: Final accuracy results per model

	Boundary Model	Entity class model 1 (meta-network category)	Entity class model 2 (meta-nw cat. + specificity)	Entity class model 3 (meta-nw cat. + subtype)	Entity class model 4 (meta-nw cat. + specificity + subtype)
Precision	93.2%	91.4%	91.9%	90.4%	90.8%
Recall	92.5%	89.7%	90.0%	88.6%	88.9%
F	92.9%	90.6%	90.9%	89.5%	89.8%
	Bound. & Class combined, rule-based	**Entity class model 1**	**Entity class model 2**	**Entity class model 3**	**Entity class model 4**
Precision	n.a.	89.7%	90.0%	88.6%	88.9%
Recall		87.7%	87.7%	86.4%	86.5%
F		88.7%	88.8%	87.5%	87.7%

3.4.8 Error Analysis

The remainder of this results section provides error analyses for the boundary model and each entity class model[11]. I decided to conduct these error analyses on the level of individual models, not the level of merged boundary and category models, in order to enable the scrutinizing of each component individually before they are fused. Also, since the combination rules used for accuracy assessment (rigorous) are not same as the ones for integrating the models into end-user software (more forgiving about false positives, details in 4), this component-wise error analysis is more insightful. For error analysis of the boundary model, I kept the outside tag in the analysis, which is a rigorous and comprehensive approach, while for the category models, I exclude the "none" category tag. The reason for this decision is that the "none" category accounts for 76.6% of all tokens in each model, which diminishes the ratio of the relevant entity classes in the ground truth, but this ratio is an important piece of information in the error analysis. However, for the previously presented assessments, the outside and none labels were treated the same as any other label since they can (and here actually do) subsume false negatives from other categories, and can produce false positives[12] and false negatives[13] themselves, which impacts the overall accuracy rate.

[11] For the boundary model and entity class models 1 and 2 I show the confusion matrices of errors in this section, for entity class models 3 and 4 those matrices are placed in the Appendix as they are very space consuming. The tables with the statistical results for the error analysis per model all share the same structure and are shown in this section. The tables and figures contain a "na" for logically not applicable attributes.

[12] False positives are entities that were detected as members of a particular class, but truly are members of a different class. Those entities are false alarms (negative interpretation) or additional, weaker suggestions that sometimes save

Several trends can be observed across all models: Differences between accuracy per class within models are much greater than differences in overall accuracy rates across models (Table 72). Within models, high accuracy is not a matter of class size (measured as the ratio of tokens in a class over the number of tokens in the corpus). This means that small as well as large classes can achieve high accuracies. Here, high means around and above the overall accuracy for a model as shown in Table 72, and low means rates below of that.). However, the inverse of this effect is not true: low accuracy rates are only obtained for small classes (excluding the "none" label for categories). In fact, for all accuracy rates below 84.5%, the size of the impacted classes is less than 2% each, and the total size of the impacted classes is less than 10% of the corpus (again, excluding the "none" label).

Table 73: Error analysis, boundary model (absolute values)

	Prediction					
Ground Truth	unigram	unigram	unigram	unigram	unigram	Sum
unigram	99,384	852	203	1,091	8,802	110,332
begin	1,049	56,964	1,461	56	2,011	61,541
inside	234	1,816	36,412	1,111	2,325	41,898
end	1,218	25	1,127	58,003	1,168	61,541
outside	5,782	1,684	1,840	1,080	890,182	900,568
Sum	107,667	61,341	41,043	61,341	904,488	1,175,880

Table 74: Error analysis, boundary model (ordered by natural sequence of an expression)

Boundary Label	**Accuracy**	**False negatives**	**False positives**	**Ratio of size**	**Tokens in ground truth**	**Correct tokens**	**False negatives**	**False positives**
unigram	90.1%	9.9%	7.7%	40.1%	110,332	99,384	10,948	8,283
begin	92.6%	7.4%	7.1%	22.4%	61,541	56,964	4,577	4,377
inside	86.9%	13.1%	11.3%	15.2%	41,898	36,412	5,486	4,631
end	94.3%	5.7%	5.4%	22.4%	61,541	58,003	3,538	3,338
outside	98.8%	1.2%	1.6%	76.6%	900,568	890,182	10,386	14,306

The more detailed the entity class models are, the larger is the number of low-performing classes. These results support my strategy of consolidating small classes prior to learning. A similar trend can be observed for the ratio of false positives and false negatives: for most of the highly accurate classes, the ratio of false positives is higher than the ratio of false negatives, while this trend flips over for low performing classes. For practical purposes, both error types are

entities from being lost to the "none" class in case they are assigned to some alternative class (positive interpretation).

[13] False negatives are entities that were not detected as members of a particular class, but actually are members of that class. Those entities are missed entries for a class.

most detrimental when false negatives are assigned to the "outside" or "none" class. This is because for the integrating the models into a software available to end users as described in section 4, all other types of error are preserved and explicitly marked. The results do not suggest any apparent relationship between class accuracy rates and the amount of false negatives that the "outside" or "none" label account for per class, and the ratio of these two labels among the false negatives can be anywhere between very small and very large.

Table 75: Error analysis, entity class model 1 (absolute values)

	Prediction											
Ground Truth	**agent**	**attribute**	**event**	**knowledge**	**location**	**none**	**organization**	**org-att**	**resource**	**task**	**time**	**Sum**
agent	45,346	10	21	103	367	2,541	988	48	80		24	49,528
attribute	7	29,847		12	7	1,581	27		208		396	32,085
event	26		533	45	13	69	21	1	4		39	751
knowledge	309	25	5	1,721	111	629	274	20	46		54	3,194
location	665	37	2	89	20,269	1,600	923	10	58		23	23,676
none	990	1,557	24	483	717	889,025	3,217	34	1,379	22	3,120	900,568
organization	2,417	76	3	296	1,205	5,298	71,623	50	150		54	81,172
org-att	116	2		14	43	79	82	4,058	12		4	4,410
resource	286	301	6	128	87	2,678	310	10	34,268		72	38,146
task	10					66	5			17		98
time	23	614	5	28	5	2,178	17	9	18	1	39,354	42,252
Sum	50,195	32,469	599	2,919	22,824	905,744	77,487	4,240	36,223	40	43,140	1,175,880

Table 76: Error analysis, entity class model 1 (sorted by decreasing accuracy)

Entity Class	**Accu-racy**	**False Nega-tives**	**False Posi-tives**	**Size of cat. in ground truth**	**Tokens in cat.**	**Accu-rate pre-dictions**	**False Nega-tives**	**False Posi-tives**
time	93.1%	6.9%	8.8%	15.3%	42,252	39,354	2,898	3,786
attribute	93.0%	7.0%	8.1%	11.7%	32,085	29,847	2,238	2,622
org-att	92.0%	8.0%	4.3%	1.6%	4,410	4,058	352	182
agent	91.6%	8.4%	9.7%	18.0%	49,528	45,346	4,182	4,849
resource	89.8%	10.2%	5.4%	13.9%	38,146	34,268	3,878	1,955
organization	88.2%	11.8%	7.6%	29.5%	81,172	71,623	9,549	5,864
location	85.6%	14.4%	11.2%	8.6%	23,676	20,269	3,407	2,555
event	71.0%	29.0%	11.0%	0.3%	751	533	218	66
knowledge	53.9%	46.1%	41.0%	1.2%	3,194	1,721	1,473	1,198
task	17.3%	82.7%	57.5%	0.0%	98	17	81	23

Across the various entity class models, we generally obtain very high accuracy rates (in the 90ies) for the categories agent, attribute and time, high rates (upper 80ies) for organizations, locations and resources, medium rates (70ies) for events, and low rates (50ies and less) for knowledge and tasks. Regardless of the model, all variations of task and knowledge are consistently ranking lowest. For locations, specific instances are predicted with higher accuracy than generic ones, and vice versa for resources.

Table 77: Error analysis, entity class model 2 (absolute values)

	Predictions																
Ground Truth	agent gen.	agent spec.	attribute na	event spec.	knowledge spec.	location gen.	location spec.	none	org. gen.	org. spec.	org-att spec.	resource gen.	resource na	resource spec.	task na	time na	Sum
agent gen.	25,221	56	6	5	28	17	33	2,151	349	96	14	5	20	4		8	28,013
agent spec.	19	19,646	5	12	137	1	482	441	6	610	15		18	101		22	21,515
attribute na	1	3	29,890	1	13	3	7	1,626	1	21			101	17		401	32,085
event spec.		19	1	540	45		19	67	1	23	3		1	1		31	751
knowledge spec.	23	183	37	8	1,750		138	648	2	295	16		21	28		45	3,194
location gen.	22	2			2	3,256	15	981	117	15			14			5	4,429
location spec.	12	388	40	2	93	18	17,456	579	4	583	15		16	22		19	19,247
none	636	207	1,486	27	571	426	343	889,749	1,021	1,668	34	204	1,041	50	30	3,075	900,568
org. gen.	462	3	14		13	93	6	1,259	17,677	70	2	10	4			3	19,616
org. spec.	104	1,214	63	7	392	1	1,111	4,014	75	54,313	59	1	40	117		45	61,556
org-att spec.	49	18	8		21		55	105	1	74	4,063		5	7		4	4,410
resource gen.	1	1	1		3	2	2	345	27	5		1,002	2	2		4	1,397
resource na	20	27	215		21	27	21	2,021	10	38	16		32,996	32		39	35,483
resource spec.	14	104	97	4	139		85	170	3	226	4	1	29	356		34	1,266
task na	1	1			2		1	61		3					29		98
time na	5	11	564	14	27	1	6	2,101	1	14	9		12	8		39,479	42,252
Sum	26,590	21,883	32,427	620	3,257	3,845	19,780	906,318	19,295	58,054	4,250	1,223	34,320	745	59	43,214	1,175,880

Table 78: Error analysis, entity class model 2 (sorted by decreasing accuracy)

Entity Class	Accu-racy	False Nega-tives	False Posi-tives	Size of cat. in ground truth	Tokens in cat.	Accu-rate pre-dictions	False Nega-tives	False Posi-tives
time na	93.4%	6.6%	8.6%	15.3%	42,252	39,479	2,773	3,735
attribute na	93.2%	6.8%	7.8%	11.7%	32,085	29,890	2,195	2,537
resource na	93.0%	7.0%	3.9%	12.9%	35,483	32,996	2,487	1,324
org-att specific	92.1%	7.9%	4.4%	1.6%	4,410	4,063	347	187
agent specific	91.3%	8.7%	10.2%	7.8%	21,515	19,646	1,869	2,237
location specific	90.7%	9.3%	11.7%	7.0%	19,247	17,456	1,791	2,324
org. generic	90.1%	9.9%	8.4%	7.1%	19,616	17,677	1,939	1,618
agent generic	90.0%	10.0%	5.1%	10.2%	28,013	25,221	2,792	1,369
organization	88.2%	11.8%	6.4%	22.4%	61,556	54,313	7,243	3,741
location generic	73.5%	26.5%	15.3%	1.6%	4,429	3,256	1,173	589
event specific	71.9%	28.1%	12.9%	0.3%	751	540	211	80

resource generic	71.7%	28.3%	18.1%	0.5%	1,397	1,002	395	221
knowledge	54.8%	45.2%	46.3%	1.2%	3,194	1,750	1,444	1,507
task na	29.6%	70.4%	50.8%	0.0%	98	29	69	30
resource specific	28.1%	71.9%	52.2%	0.5%	1,266	356	910	389

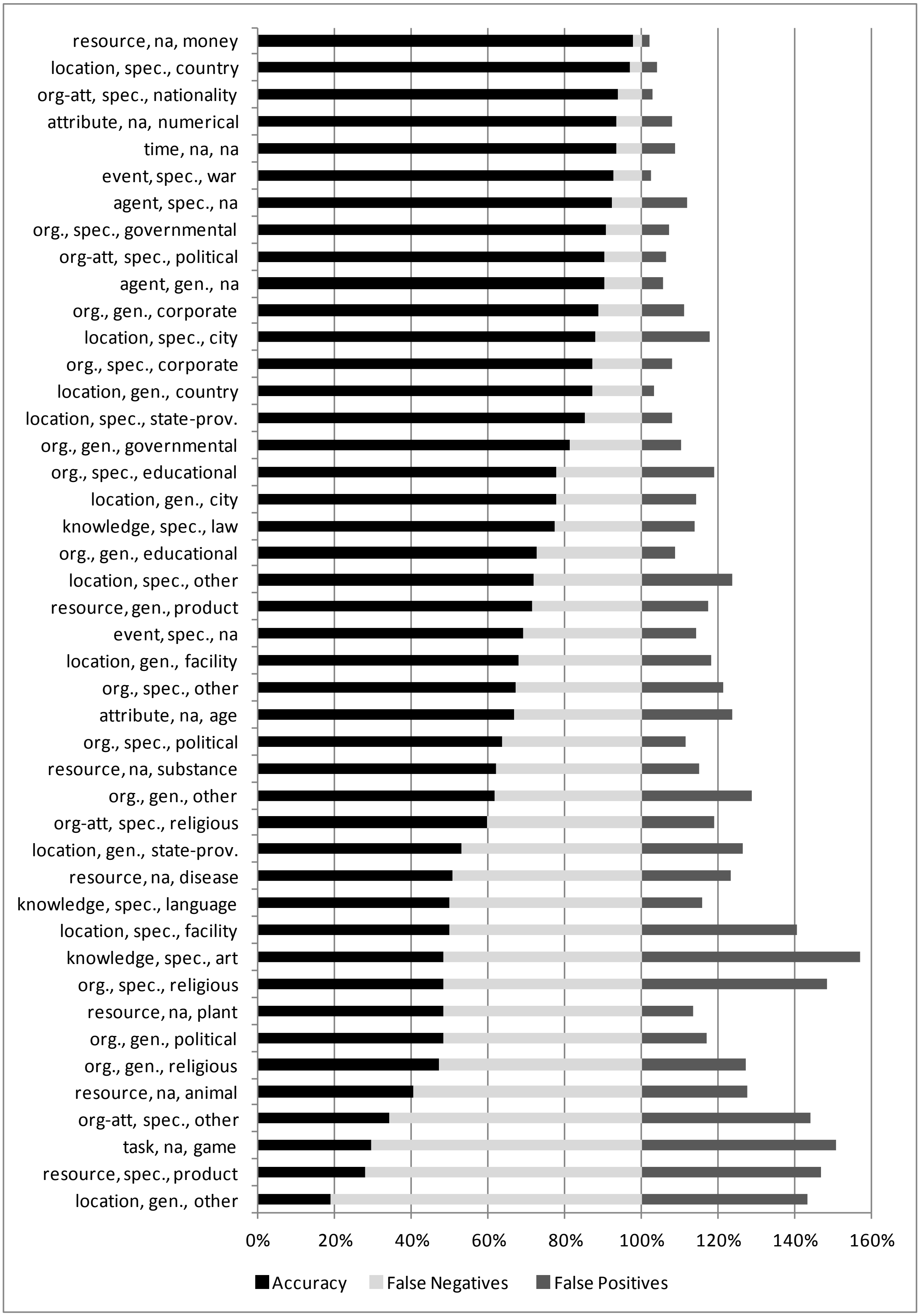
Table 79: Error analysis, entity class model 3 (sorted by decreasing accuracy)

Entity Class	Accuracy	False Negatives	False Positives	Size of cat. in ground truth	Tokens in cat.	Accurate predictions	False Negatives	False Positives
resource money	97.5%	2.5%	2.1%	11.5%	31,686	30,905	781	647
location country	94.4%	5.6%	4.9%	2.4%	6,701	6,329	372	326
attribute numerical	93.6%	6.4%	8.2%	11.3%	30,991	28,995	1,996	2,598
time na	93.3%	6.7%	8.7%	15.3%	42,252	39,439	2,813	3,760
org-att nationality	93.3%	6.7%	4.4%	1.3%	3,538	3,300	238	151
agent na	91.7%	8.3%	9.9%	18.0%	49,528	45,418	4,110	4,987
event war	90.2%	9.8%	2.7%	0.0%	122	110	12	3
organization gov.	88.7%	11.3%	8.5%	4.0%	10,925	9,691	1,234	906
org-att political	88.1%	11.9%	9.5%	0.2%	682	601	81	63
org. corporate	86.3%	13.7%	9.5%	23.0%	63,382	54,724	8,658	5,742
location city	84.5%	15.5%	17.9%	2.9%	7,889	6,667	1,222	1,450
location state-prov	80.4%	19.6%	9.7%	1.3%	3,530	2,838	692	304
organization edu	77.9%	22.1%	13.6%	0.5%	1,246	971	275	153
knowledge law	76.6%	23.4%	11.4%	0.3%	907	695	212	89
location other	70.8%	29.2%	26.2%	0.8%	2,083	1,475	608	523
attribute age	69.8%	30.2%	21.6%	0.4%	1,094	764	330	210
event na	67.7%	32.3%	16.5%	0.2%	629	426	203	84
organization other	65.9%	34.1%	21.0%	1.7%	4,669	3,077	1,592	819
organization political	63.2%	36.8%	9.7%	0.3%	798	504	294	54
location facility	62.8%	37.2%	21.8%	1.3%	3,473	2,182	1,291	610
resource substance	60.4%	39.6%	14.2%	1.0%	2,808	1,697	1,111	281
org-att religious	59.6%	40.4%	21.1%	0.0%	94	56	38	15
resource disease	51.3%	48.7%	17.4%	0.1%	378	194	184	41
organization religious	50.7%	49.3%	34.2%	0.1%	152	77	75	40
resource product	50.1%	49.9%	23.6%	1.0%	2,663	1,334	1,329	412
knowledge language	50.0%	50.0%	8.5%	0.0%	86	43	43	4
resource plant	48.5%	51.5%	12.7%	0.1%	198	96	102	14
knowledge art	47.3%	52.7%	58.6%	0.8%	2,201	1,040	1,161	1,473
resource animal	40.7%	59.3%	24.7%	0.2%	413	168	245	55
org-att other	34.4%	65.6%	35.3%	0.0%	96	33	63	18
task game	24.5%	75.5%	52.0%	0.0%	98	24	74	26

Table 80: Error analysis, entity class model 4 (sorted by decreasing accuracy)

Entity Class	Accuracy	False Negatives	False Positives	Size of cat. in ground truth	Tokens in cat.	Accurate predictions	False Negatives	False Positives
resource, na, money	97.7%	2.3%	2.1%	11.5%	31686	30958	728	662
loc., spec., country	97.0%	3.0%	4.1%	2.1%	5708	5538	170	234
org-att, spec., nat.	93.8%	6.2%	2.9%	1.3%	3538	3319	219	100
attrib., na, numerical	93.4%	6.6%	8.2%	11.3%	30991	28960	2031	2580
time, na, na	93.4%	6.6%	8.7%	15.3%	42252	39464	2788	3772
event, spec., war	92.6%	7.4%	2.6%	0.0%	122	113	9	3
agent, spec., na	92.3%	7.7%	11.8%	7.8%	21515	19849	1666	2649
org., spec., gov.	90.8%	9.2%	7.3%	3.1%	8404	7629	775	597
org-att, spec., pol.	90.5%	9.5%	6.5%	0.2%	682	617	65	43
agent, gen., na	90.2%	9.8%	5.8%	10.2%	28013	25263	2750	1562
org., gen., corporate	88.7%	11.3%	11.1%	5.6%	15305	13581	1724	1691
loc., spec., city	88.1%	11.9%	18.0%	2.7%	7512	6615	897	1452
org., spec., corporate	87.2%	12.8%	8.0%	17.5%	48077	41938	6139	3651
loc., gen., country	87.1%	12.9%	3.5%	0.4%	993	865	128	31
loc., spec., state-prov.	85.4%	14.6%	8.1%	1.1%	3133	2675	458	237
org., gen., gov.	81.4%	18.6%	10.4%	0.9%	2521	2051	470	237
org., spec., edu.	77.8%	22.2%	19.2%	0.4%	1001	779	222	185
loc., gen., city	77.7%	22.3%	14.3%	0.1%	377	293	84	49
knowledge, spec., law	77.5%	22.5%	13.8%	0.3%	907	703	204	113
org., gen., edu.	72.7%	27.3%	8.7%	0.1%	245	178	67	17
loc., spec., other	71.8%	28.2%	23.7%	0.7%	2014	1447	567	450
res., gen., product	71.7%	28.3%	17.5%	0.5%	1397	1001	396	213
event, spec., na	69.0%	31.0%	14.4%	0.2%	629	434	195	73
loc., gen., facility	67.9%	32.1%	18.3%	0.9%	2593	1760	833	395
org., spec., other	67.1%	32.9%	21.2%	1.2%	3326	2233	1093	600
attribute, na, age	66.9%	33.1%	23.8%	0.4%	1094	732	362	228
org., spec., political	63.8%	36.2%	11.4%	0.2%	647	413	234	53
res., na, substance	62.0%	38.0%	14.9%	1.0%	2808	1742	1066	306
org., gen., other	61.6%	38.4%	28.8%	0.5%	1343	827	516	334
org-att, spec., religious	59.6%	40.4%	18.8%	0.0%	94	56	38	13
loc., gen., state-prov.	52.9%	47.1%	26.6%	0.1%	397	210	187	76
resource, na, disease	50.8%	49.2%	23.5%	0.1%	378	192	186	59
know., spec., language	50.0%	50.0%	15.7%	0.0%	86	43	43	8
loc., spec., facility	49.8%	50.2%	40.7%	0.3%	880	438	442	301
knowledge, spec., art	48.5%	51.5%	57.1%	0.8%	2201	1068	1133	1422
org., spec., religious	48.5%	51.5%	48.4%	0.0%	101	49	52	46
resource, na, plant	48.5%	51.5%	13.5%	0.1%	198	96	102	15
org., gen., political	48.3%	51.7%	17.0%	0.1%	151	73	78	15
org., gen., religious	47.1%	52.9%	27.3%	0.0%	51	24	27	9
resource, na, animal	40.4%	59.6%	27.7%	0.2%	413	167	246	64
org-att, spec., other	34.4%	65.6%	44.1%	0.0%	96	33	63	26
task, na, game	29.6%	70.4%	50.8%	0.0%	98	29	69	30
res., spec., product	28.0%	72.0%	47.0%	0.5%	1266	354	912	314
loc., generic, other	18.8%	81.2%	43.5%	0.0%	69	13	56	10

Figure 11: Error analysis, class model 4



3.4.9 Integration of prediction models into end-user software

Once the accuracy of the final models had been evaluated, the remaining task for this project is to make the models publically available in a software product. The goal with this step is to provide this prediction technology such that people from different backgrounds with potentially very little expertise in natural language processing can use it for their text analysis projects. The integration process is described in detail in chpater 4.1 in the operational chapter.

3.5 Limitations

The prediction capabilities of the built model strongly depend on the training data. Even though I chose a training dataset with a large number of examples and a suitable set of categories and category attributes, there are several limitations with the BBN dataset: First, the data are from a single source, namely the Wall Street Journal. Second, the data represent a single genre and well defined domain, i.e. newspaper articles. Thus, the models can be expected to generalize with less accuracy to different genres and writing styles than to the training domain. Third, the articles are from 1989, which implies that terms and phrases might be outdated, and many agents and other entities that are relevant today might not occur in the data. This issue might already have been mitigated to some degree by using a lookup dictionary that is based on current news data. Fourth, since the learning data is in English only, the resulting models cannot be expected to generalize to other languages. Fifth, BBN contains only a few types of activities, which limits our ability to predict task and events of the type that the meta-network model expects. Sixth, the data contained various inconsistency issues as outlined in section 3.3.1 that we corrected for as we found them prior to learning. However, when evaluating the results, we saw that a handful of entities in the marked up files crossed line breaks or paragraph breaks in a way that a multi-word expressions are interspersed with a few additional spaces, e.g. "Cie.   Fianciere de Paribas". The learner has picked up on these few problematic cases and developed some reasoning about them. While these cases are noisy and could impact the accuracy of the overall model, they might reflect scenarios that can be found in new data as well. Overall, the outlined limitations can be addressed by enhancing the learned models or building new models by learning with more recent data that originates from more sources, covers more domains, and contains more examples of activities.

Including other feature types, using a different combination of feature types, or applying a different iteration rate might all have led to better and potentially more accurate or more robust prediction models. The part of speech tagger that was used as a feature type for this project is not error free to begin with, but achieves about 93% accuracy. This issue represents a general limitation with features that require pre-processing of the text data: the pre-processing routines are imperfect in terms of their accuracy. As a result, errors with these routines get propagated throughout the learning process. Furthermore, generating these features further increases the runtime costs (Sarawagi, 2008).

The built models retrieve entities based on an ontology that originates from social science research, namely the meta-network model. This model is designed to capture the who, what, why, where and how of events. However, for constructing network data that adhere to alternative classification schemas, different ontologies and respective models might be necessary.

Finally, training models with CRF has high run time costs. For example, building the final class label prediction models that outputs a meta-network category along with a specificity attribute and a category subtype per entity took nine days. This time constraint requires careful planning of experiments for testing the impact of features on prediction accuracy. Such experimentation is further complicated by the fact that small iterations rates (in the case of this study less than 300) do not necessarily allow for extrapolating to results with higher, more appropriate iterations rates. However, once the models have been built, applying them for inference to new data is speedy, as demonstrated in the next chapter.

3.6 Conclusions and Future Work

Two main contributions have been made with this project: first, I have developed a highly accurate computational solution to the extraction of entities from text data. The approach I used for building these prediction models is interdisciplinary in that it combines a theoretically grounded model from organization science for informing the definition of relevant entity classes with cutting edge methods from natural language processing and machine learning. The obtained accuracy rates are on a par with rates from alternative, top-performing entity extractors. However, beating benchmarks was not the goal here. Rather, the objective was to build an entity extractor that end-users can apply in the process of constructing one-mode and multi-mode network data that support them in answering substantive question about socio-technical networks. Delivering such a product as part of a publically available tool (AutoMap) is the second contribution with this project. Going from learned models to usable technology involved its own challenges. An example is the designing of rules for handling false positives such that end-users are best supported in their needs, which required different rules than the ones I applied for the rigorous assessment of the accuracy of the learned models.

At the beginning of this chapter I had defined several sub-goals for this project. Table 81 summarizes how they have been met, and points out the practical relevance of these objectives.

Table 81: How project goals have been met and practical relevance of solutions

Goal	Delivered outcome	Practical relevance
1. Automation	- Scalable and publically available solution to entity extraction.	- Supports analysis of large text datasets. - Reduces time and labor costs for thesaurus construction.
2. Abstraction of terms to concepts or higher level aggregates	- Text level terms are associated with meta-network categories that encode different levels of detail, namely a specificity value and/ or a subtype per entity. Since prediction results might differ between reducing a complex model to a simpler model and training a simpler model separately, models at five different levels of granularity were built and evaluated.	- Allows user to choose the level of granularity the best fits their needs. - Allows user to balance accuracy and granularity based on their needs.
3. Generalization	- Ability to identify new and unseen instances of entity classes and entity attributes.	- Faster analysis of and adaption to new corpora. - Reduced time and labor costs for thesaurus construction.
4. Support users in addressing substantive and meaningful questions about socio-technical networks	- Ability to extract meta-network data from texts. These data can be further analyzed in ORA, which provides metrics defined over non-generic entity classes.	- Move beyond the extraction and analysis of social networks (agent by agent connection) or generic one-mode networks to the analysis of multi-mode, socio-technical networks.
5. N-gram detection	- Correctly identify boundary and class of multi-word entities.	- The boundary class models that facilities the detection of entities (unigrams and multi-word expressions) is particular useful for constructing one-mode networks and content analysis. Once these entities are identified, they can also be classified, which supports the construction of multi-mode networks.
6. Allow terms to belong to multiple entity classes instead of just one.	- Ability to assign identically spelled terms to multiple meta-network categories. - Differentiate terms based on predicted label and for the NORP class also on part of speech.	- Contributes to the disambiguation of homonyms. . - Reduced loss of relevant information over current thesaurus creation technique in AutoMap.
7. Entity Extraction (as opposed to focus on Named Entity Extraction)	- Ability to extract entities that are a) referred to by a name or not and b) instances of classes where many entities are not named.	- Allows for distinguishing between generic and specific entities, which is particularly useful when term presenting roles of social agents subsume a large number of references.

From a NLP perspective, the findings from this study imply several conclusions about the impact of engineering decisions and particular features types on the accuracy and required training as summarized in Table 82. The most unexpected finding was that large differences in model complexity (number of prediction classes, which impacts the number of states and edges in the probabilistic graphical model) lead to only small differences in accuracy rates. In contrast to my hypothesis, less complex models are not necessarily more accurate than more complex ones. With respect to the per class accuracy within prediction models, the results indicate that high accuracy is not a matter of class size, but low accuracy was only observed for small classes. Considering both findings together leads to the following recommendation for designing entity extractors: it is critical to find a good balance between consolidating small class into larger aggregates and avoiding the fusion of classes with very different (weights per) features, which potentially dilutes the expressiveness of features.

Table 82: Impact of variable on outcomes

Variable	Accuracy	Training Time
Baseline	large	small
Syntax Features (POS)	small	small
Lexical Features (Dictionary, hard match)	large	small
Iteration Rate	large	large
Complexity of Category Schema/ Model	small	large

With respect to feature types, in my results the part of speech tags were the weakest contributor to accuracy. This could be due to the fact that part of speech tags are not orthogonal to other clues, or that other syntax features might be more appropriate. In future work, it seems worthwhile to test more advanced syntactic features, such as the constituent of a parsing tree that per token. Also, the results show that it is important to test the isolated impact of each baseline feature as gains from eliminating non-contributing features can be substantial.

When the goal is to provide the entity extractor to end-users, it is furthermore crucial to test if the models that the learning system outputs are readily usable for inference in another environment. In the case of this study, adjustments were needed that had to be represented in the learning output directly and thus required retraining of the models after these discrepancies were detected. To harness those situations, I recommend plugging in a first output model, e.g. one from learning with the feature baseline only, into the external inference environment in order to identify any necessary adjustments. This eliminates time for retraining when it comes to building the final models with the best and most robust feature set found.

The presented solution involves several considerations that are particular to the goal of aiming for practical usefulness of the models, and are fairly independent from the NLP and machine learning methods part: the models were built such that they are particularly suitable for extracting relevant entities from documents about socio-technical systems. One strategy for achieving this goal was to use a theoretically grounded model from organizations science to inform the selection of relevant entity classes. Furthermore, the generated models support the consideration of entity classes where many instances are common nouns and noun phrases, e.g. in the resource class. Specific and generic entities, which often means entities that are referred to be a name or not, are distinguished from each other. This is important for keeping roles versus specific references to agents separate from each other. Finally, I have designed and implemented the way that outputs are generated from these models such that the output data include entities for which a non-outside boundary label has been found but no class label and vice versa, or for which other discrepancies between both labels exist. For assessing the accuracy of prediction models, these cases were handled differently, i.e. more rigorously as defined by standard information extraction assessment procedures. There, such conflicting cases are considered as inaccurate and are disregarded from final outputs. However, for practical applications of parsing entities from news wire data and other accounts of event coverage, optimizing on error reduction might be less important than retrieving the largest possible set of potentially relevant entities. The presented solution implies the assumption that end-users might be willing to comprise some accuracy in label assignment (precision) for a greater coverage of retrieved entities (recall) for two reasons: First, entirely rejected entities might be hard to retrieve otherwise. Second, finding a class for yet unlabeled but retrieved entities or correcting the class of entities for which discrepancies are explicitly marked as such might be more acceptable than knowing that those cases are returned altogether.

The lowest performing classes in the models I built are activities in general (tasks and events), as well as knowledge and specific resources. In future work, these limitations can be addressed by using additional learning data that contains more examples for these classes, and by only merging classes that are similar in content as well as (weights of) features. For this project, category merging was driven by resembling the categories in the meta-matrix model and avoiding overly small classes. Furthermore, the learning data for this project was from a single, somewhat dated source and genre. In order to provide more flexible models with a potentially higher capacity to provide correct predictions for corpora that feature more current style and content, we should also consider more recent training data from multiple domains and genres.

4 From Experimental Results to Practical Applications

This thesis does not stop at providing experimental results for the rigorous evaluation of the impact of relation extraction sub-routines on network data and analysis results (chapter 2) and the construction of an entity extractor (chapter 3), but also reports on the transition from providing these results to understanding their practical implications (this chapter). This chapter makes a contribution to the practical usefulness of methods for relation extraction from text data by developing answers to the following question:

- What steps are necessary for making the outcome of the experiments and evaluation studies that are based on ground-truth data operational?
- What challenges and limitations apply when brining the understanding about these experimental results into application contexts?

This chapter is structured as follows: in section 4.1, practical implication and respective recommendation for relation extraction are developed based on the outcome of the study of the impact of coding choices and network data and analysis results. In section 4.2, I describe the steps, respective issues and developed solutions for making the entity prediction models available as a new functionality in an existing end-user product.

4.1 Impact of Coding Choices for Reference Resolution and Windowing on Network Data and Analysis Results: Implications and Recommendations for Applied Work

The results for the impact of reference resolution on network data greatly differ depending on the chosen approach for normalizing nodes: if node IDs are available that reflect the true identify of nodes, I recommend working with these IDs instead of using node names as proxies for node IDs, which implies the risk of merging different nodes with identical surface forms. The ORA software supports this strategy by allowing for node ID's that are different than the node names. For example, homonyms can be disambiguated by different node IDs. If no such node IDs are available, which is often the case for networks extracted from texts, and nodes are disambiguated and consolidated based on their spelling, conducting any reference resolution technique is not necessarily worthwhile with respect to key player analyses and the majority of graph-level network measures because the results will still be strongly distorted. If this approach is chosen, the obtained results will not resemble findings that would be obtained by using ground truth data. To prevent his outcome under the condition that no node IDs are available, I recommend not to conflate nodes based on their spelling, but trying to perform node disambiguation and consolidation as well as possible. The following strategies can be used to this effect:

- After important raw text data into a text analysis tool and prior to performing reference resolution, the following techniques can be used; all of which are available in AutoMap:
    - Disambiguate entities based on their part of speech (Diesner & Carley, 2008b).
    - Identify meaningful multi-word expressions such that some individual tokens become part of distinct units.
    - Identify the node class of entities, and disambiguate nodes and multi-word expressions based on the node class.

The entity extraction models that were developed in the previous chapter can help with all three of these steps. Therefore, the entity extractor built for this thesis not only serves the purpose of identifying nodes for the construction of network data, but also facilitates pre-processing steps that are crucial for relation extraction.

If the resources for performing reference resolution are limited, I further recommend focusing on co-reference resolution rather than anaphora resolution because it has a bigger impact on the network data level. This decision further requires sticking with key player analysis instead of calculating network metrics when analyzing the network data.

When it comes to selecting a reference resolution tool or technique, differences in accuracy do matter, especially if the harmonic mean of recall and precision is below 90%. Therefore, I recommend looking for a tool that achieves the best accuracy for a given domain or genre.

When connecting nodes into edges, caution is needed if windowing is chosen as the link formation mechanism. This is because the rate of false positives can be very high: according to the findings from this thesis, nine out of ten links can be false positives at a decent window size across time, various domains, and writing styles. To lower this risk, the following strategies can be applied, e.g. in AutoMap:

- Code roles and node attributes not as actual node class, but as node attributes. A solution to this strategy is developed in the next chapter.
- Disregard overly common nodes for entity extraction. These nodes can be identified, for example, by (weighted) term frequency metrics computed on the entities (Diesner & Carley, 2004; Yang & Pedersen, 1997).

Based on the empirical results for the impact of proximity-based link formation on network data and analysis results, the following recommendations can be made:

- If a corpus contains an indistinguishable mixture of syntactic and semantic link, at least 90% of all links are covered with a window size of seven. Syntactic links are natural by-production of language production rules, such as links between adjectives and the proper nouns they modify. Semantic relationships are more independent from language production rules and can be orthogonal to these rules, such as the description of the type of social relationship between two agents in text data.

- If syntactically motivated links are disregarded, more than 90% of true links are typically found when using a window size of twelve. This result is robust cross genres, types of semantic relationship, and node classes.
- When using windowing as a link formation method, one needs to keep in mind that the amount of false positive links can be enormous. Again, this risk can be mitigated by coding attributes of nodes, such as roles and titles, as properties of the respective nodes instead of as separate node classes.

4.2 From Learned Models to Usable Technology: Integration of Prediction Models into End-User Software

Once the accuracy of the prediction models for entity extracted has been evaluated (as done in chapter 3), the remaining task for that project was to make the models publically available in a ready-to-use software product. The goal with this step is to provide these prediction models such that people from different backgrounds with potentially very little expertise in natural language processing can use such a technology for their text analysis projects.

The five different prediction models that have been constructed as reported on in the previous chapter were integrated into AutoMap as described in the remainder of this section. For each model, the expected accuracy rate, level of detail encoded in the models or what exactly gets predicted and an example are provided Table 83.

Table 83: Prediction models provided in AutoMap

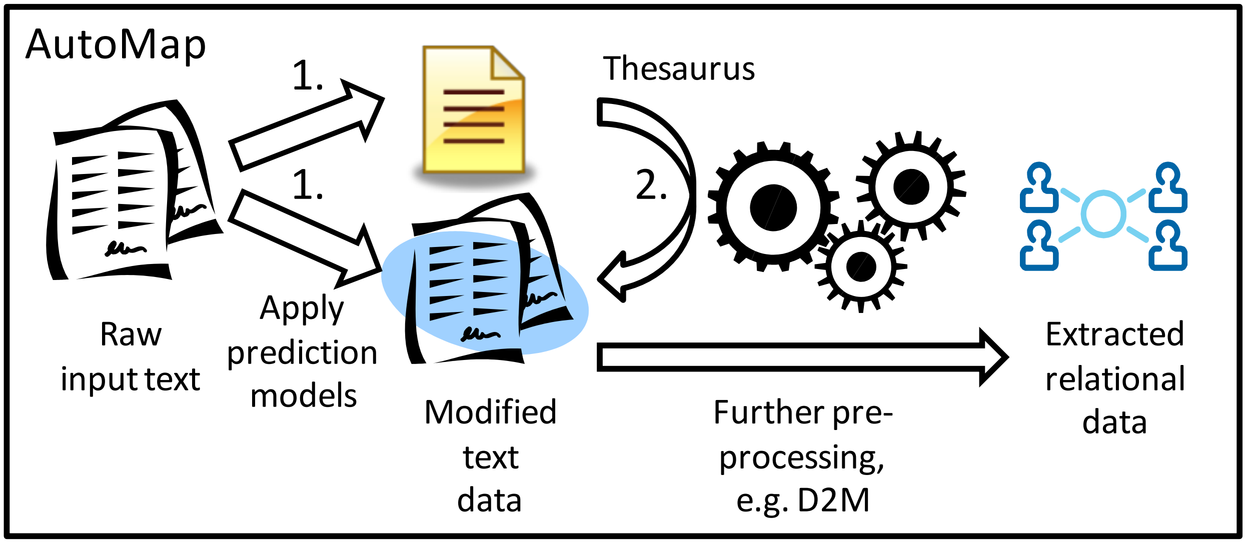
Model Name	Expected Accuracy	Boundary	Meta-network class	Specificity	Subtype	Example
Boundary model	92.9%	x				Madeleine Albright
Entity class model 1	887.%	x	x			agent
Entity class model 2	88.8%	x	x	x		agent, specific
Entity class model 3	87.5%	x	x		x	agent, political
Entity class model 4	87.7%	x	x	x	x	agent, specific, political

The intended workflow of using these models is depicted in Figure 12. The models are executed by going to the main AutoMap GUI, MenuBar, and selecting “Generate”, “Thesaurus suggestion”. Each of these models is expected to be applied to raw text data. This means that no normalization and pre-processing routines, such as reference resolution, stemming or the removal of stop words, should be applied prior to model application. Each model can be called individually. Users can run as many models as they wish in any arbitrary order; there is no interaction between the models. The output from running any one of these models is a thesaurus that gets stored at directory of the user's choice. This eliminates or reduces the need to construct thesauri by employing alternative NLP routines as described in section 5.2.2.1, which is considerably more time consuming and requires computer-supported work by humans. The output from each of the five models contains the extracted entities (one per row) and the following information per entity in tabular (comma separated values, one value per column):

- A concatenation of multi-word expressions into a single token via underscores, e.g. United Nations into United_Nations. This helps to keep entities together when they appear as nodes in a network, and complies with the standard node formatting conventions for AutoMap.
- The meta-network category per entity (for the boundary prediction model, the default class "knowledge" is used).
- Depending on the chosen prediction model, zero, one or two attributes per entity that represent the specificity and/ or subtype value if applicable. Specificity can take the values "specific", "generic", or "not applicable". In the latter case, no attribute gets output. For a list of the possible subtype values see Table 59.
- The part of speech of each token in an entity, i.e. multiple part of speech in the case of multi-word expressions.
- The cumulative frequency per entity as inferred from the text data.

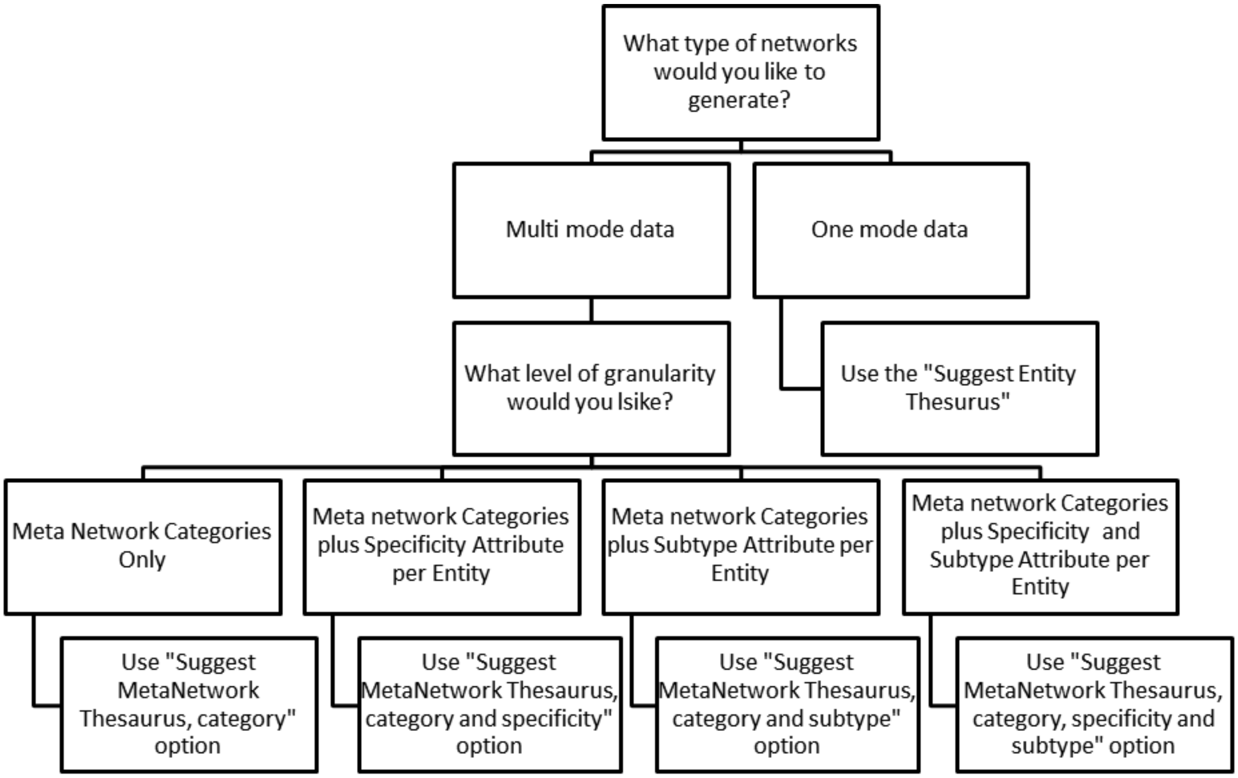
An entity's frequency is increased when two entities agree in spelling including capitalization, as well as in meta-network category, any attribute per category, and part of speech. This rule helps to disambiguate entities based on their part of speech, which is another new functionality in AutoMap that got added as part of this project. Also, this rule also supports the consolidation of entities that differ in capitalization only during thesaurus application. This could for instance apply to entities that typically occur in lower case, e.g. "apple" (the common noun), but are capitalized at the beginning of a sentence, and are still different from words that are orthographically the same, but have a different meaning (such as "Apple" as the company). I defined these rules for disambiguation and consolidation in order to prevent the loss of information that we had previously experienced but not handled in AutoMap.

Figure 12: Workflow of using prediction models in AutoMap



Users might ask themselves which model to use. To address this question, I designed a decision tree and added a visual representation of it to the prediction model routine in AutoMap (Figure 13). We also decided to make class model 2 the recommended default model in AutoMap because a) this model achieved the highest accuracy during formal evaluation and b) was at the level of detail that we needed for many of the application scenarios in the CASOS lab over the last years. The boundary prediction model is the only model that theoretically extracts uncategorized entities, which can be unigrams or multi-word expressions. These entities can be used for conducting content analysis, or as nodes for constructing one-mode networks. In the thesauri generated by applying the boundary model, the extracted entities are actually assigned to the "knowledge class" because this class is considered as the default class for text coding according to the meta-matrix model. The outputs from all of the prediction models can be used to manually consolidate synonymous entities that have different surface forms. This a form of co-reference resolution and helps to alleviate the issues with disambiguating and consolidating nodes based on spelling as identified in the previous chapter.

Figure 13: Decision Tree for prediction model selection in AutoMap


What type of networks would you like to generate?
Multi mode data
One mode data
What level of granularity would you lsike?
Use the "Suggest Entity Thesurus"
Meta Network Categories Only
Meta network Categories plus Specificity Attribute per Entity
Meta network Categories plus Subtype Attribute per Entity
Meta network Categories plus Specificity and Subtype Attribute per Entity
Use "Suggest MetaNetwork Thesaurus, category" option
Use "Suggest MetaNetwork Thesaurus, category and specificity" option
Use "Suggest MetaNetwork Thesaurus, category and subtype" option
Use "Suggest MetaNetwork Thesaurus, category, specificity and subtype" option

Next, I describe the types of challenges (marked in italics at the beginning of paragraphs) that can occur throughout this process and say how I addressed these issue for the case of AutoMap. I argue that many of these challenges generalize to a) providing an entity prediction technology as a stand-alone, end-user product or b) integrating an entity extractor into an existing systems with given constraints.

1. Training of models: For end-user applications, each model needed to be trained with all training folds and no holdout folds. I used the same feature configuration as I did for the final round of accuracy assessment (Table 71). The upper bound on training time is constrained by the most complex model (class model 4), which took about ten days to complete.

2. Separate inference engine: Integrating the models into an existing software product required the construction of an inference engine that uses outputs from the learning process (details below) to make predictions on new and unseen text data, and added this inference engine to AutoMap. This engine reuses parts of the learning code, but also required new code. The outputs from learning that needed to be migrated into AutoMap are a model file (number of features and weight per feature), a features file (each feature and its ID), and a coding files that associates numeric values of prediction classes with logical values of those classes (details on that in the next paragraph).

3. Different inference systems: AutoMap features a GUI version and a script version. While these two versions share some code, integration had to be done for each version individually. Therefore, every step described in this section was performed and validated for the GUI version and the script version separately while making sure that both versions produce identical results for this routine.

4. Incomplete learning output representation: When I integrated the first set of models into AutoMap, both, the retrieved entities and their classifications, seemed highly inaccurate. Investigating this issue revealed a critical difference between the models as they are held in memory after training and prior to evaluation, and the models that get stored out to disk. This difference is specific to the CRF technology I adopted for this project, but might generalize to other CRF implementations: when the models are temporarily stored in memory, they also keep the information about which numerical value for each class label (boundary and category) maps to which logical value for each of these labels. The CRF implementation picks these numerical values internally, implicitly and to the best of my knowledge in random order. This procedure applies not only to the boundary and category labels, but also to the features. Since I added new features to the CRF baseline, there were also numerical values for each part of speech tag and each entry in the lookup dictionary. The problem here is that once the models are stored out, this mapping is not output by default or represented in any output file. Thus, I had to re-engineer this mapping if I wanted to make my models work. However, I could not find any apparent logic, regularities, or systematic way according to which this mapping or assignment of numerical values to labels happens. Therefore, I had to retrain all models with the exact same features such that the outputs include this mapping. This retraining had no impact on model accuracy; the only difference was that the output files contained the needed mapping information.

5. Routine incompatibility: After the previous change had been made, the resulting models showed greatly improved prediction results. Nevertheless, the results still seemed less accurate than what the final results from the k-5 cross validation led me to reasonably expect. This could be due to poor generalization capabilities of the models or technical issues with integrating the models into AutoMap. Exploring this issue further first revealed a problem that might generally apply to situation in which new routines are plugged into existing, larger systems, and where the new routine reuses available functionalities. In this particular case, this existing routine was the part of speech tagger. The change regarding the tags for tokens involving digits did conflict with the POS implementation and tag set already available in AutoMap. I solved this issue by adding the part of speech tagger that I had added to the CRF environment into AutoMap. The difference between both taggers is small, but makes a big difference for the accuracy of prediction models.

6. Input representation issues: At this point, the prediction quality of the models still seemed lower than what I expected, and I was still hoping that this drop in performance was not due to the quality or generalizability of the models themselves, but the way they were integrated into AutoMap. The next issues that I identified were differences between how input data are represented in AutoMap versus how the learning data were formatted. In order to solve this problem, I went back to the BBN data and identified these formatting particularities by carefully going through the data and paying special attention to non-letter, non-digit characters. Next, I adjusted the formatting of the texts that the prediction models in AutoMap take as an input such that they resemble the following idiosyncrasies: in BBN, sentence marks are space-separated from the last word in a sentence, while other dots, such as in Mr. or U.S., are not space-separated from the tokens they belong to. I reused the sentence splitter that I had previously integrated into AutoMap for the purpose of determining sentence boundaries and distinguishing them from other dots (Diesner & Carley, 2004). Also, in BBN, commas have a space character right and left from them, and the same is true for various other non-digit, non-letter symbols, e.g. hyphens and percentage signs. However, there are exceptions to this rule, e.g. dashes within multi-word units, such as in "money-market". Finally, genitive markers of nouns, e.g. "parent 's", and negations of verbs that are part of the word, such as "did n't" or "is n't", are space-separated from the main verb as shown in the examples above. Once those changes were made, the prediction accuracy of the models in AutoMap was improved and seemed satisfying. There are two ways to realize these changes: they could be represented only internally, or the adjusted formatting could be displayed to the user as well. Since one of the main purposes with these models is to generate thesauri that users can apply to the text data when generating networks data, it is crucial that the entities in the prediction outputs match the text data. Thus, I decided to store the modified text data so that users can load them for further work and thesaurus application if they wish.

7. Trading off conciseness and certainty for recall: Next, additional changes were necessary to ensure that the new prediction routines support end-users in addressing substantive questions about socio-technical networks. First, I adjusted the rule set for combining the boundary and category model (to the boundary dominating policy) such that fewer entities are missed than with the rigorous rule set used for model assessment up to here. During error analysis, I observed that oftentimes, the boundary label is correctly indicating an entity and a class label is suggested as well, but the category prediction is not perfectly accurate and rather returns a reasonable alternative. For example, "consultants" were predicted as a generic organization, but the ground truth labels them as a generic agent. For the end user, such false positives might still be relevant: for practical applications of entity extraction, recall is often considered more important than precision (Sarawagi, 2008). This is because incorrect class labels can be corrected for by hand, but entities that are not returned as a potentially relevant hit at all would be hard to retrieve otherwise. Therefore, the modified combination rules for the end-user tool penalize the following discrepancies less for inference than for training: tokens with a non-outside boundary label but no class label as well as the inverse case are both output and are explicitly marked as potentially useful additional hits. These tokens might be false positives or true negatives. Except for these changes, the same combination rules as described above are applied.

8. Category adjustment: Finally, BBN contains four categories of the NORP type (nationality, other, religion, political, for details see Table 59). Instances of NORP are either specific agents or organizations or attributes. Since end-users might want to be able to distinguish between these cases, I separate them for application in AutoMap based on their part of speech after checking the hits that these classes returns: all instances that are labeled as nouns (NN, NNP, NNS, NNPS) or personal pronouns are categorized as specific organizations of the respective subtype (if applicable in the entity class model). All other instances are assigned to the attribute category.

9. Output representation issues: A naturally suitable output format for the entity lists or thesauri generated by the prediction models would be a tab delimited format. However, in AutoMap, these types of output have to be in csv format. The problem here is that retrieved entities may contain commas, which would mess the csv outputs. Note that these outputs are used for further computations and thus have to adhere to certain constraints as given by the AutoMap and ORA toolkits. In order to accommodate the change from tab delimited (initial output) to csv, I added a functionality that removes commas from the text data after prediction and string the modified text data along with the prediction outputs at a separate, user-defined location. *10. Usability:* Since the proper application of these models in AutoMap (or anywhere else) is not necessarily intuitive to end-users, different types of documentation are needed. In order to assist users in selecting the model that best fits their needs, I added a decision tree that differentiates the models based on the level of detail they encode and their accuracies (Figure 13). Also, I wrote a user's guide for this sub-routine that is part of the AutoMap help system.

11. Reusability: Finally, I built the learning technology for this project such that it can be re-used by CASOS members to train models that are based on modified or different ontologies, or use different features.

In summary, integrating the learned models into an existing software product implies additional tasks and challenges that are not necessarily foreseeable during the model construction stage and might even require the re-training of the models. Overall, the time costs for making the learned models publically available in a ready-to-use fashion are significant: the described integration process took about as long as selecting features and training and testing the models took all together.

5 Comparison of Relation Extraction from Texts including Entity Extraction to Alternative Methods for Network Data Construction in Application Contexts

In this chapter, I demonstrate the end-to-end process of going from raw text corpora to network data to analysis results. This chapter puts the knowledge about of the impact of coding choices on network analysis (chapter 2) and the entity extractor (developed in chapter 3, integrated into end-user product in section 4.2) into real-world application contexts; addressing two questions:

- How do prediction models for entity extraction perform beyond lab studies, i.e. k-fold cross validation, when used for real-world applications?

In chapter 3, the performance and accuracy of the prediction models was evaluated via k-fold cross validation on held-out portions of the corpus used for training the models. In this chapter, I further evaluate these models by applying them to new and unseen (with respect to the models) text data that differ in genre, domain and publication data from the training data and among each other. This validation enables us to better understand how the constructed prediction models generalize.

However, relation extraction by using the prediction models for entity detection is only one way for coding texts as networks. In order to put this approach into the wider context of relation extraction methods, this chapter addresses a complementary second question:

- How do the network data and respective analysis results obtained by using the prediction models as part of the relation extraction process compare to alternative methods for constructing network data from the same corpora?

Answers to this question help us to understand how the relation extraction method supported by the entity extractor developed and validated herein generalizes, i.e. compares to alternative methods.

5.1 Motivation and Research Questions

The formal evaluation of the prediction models (chapter 3) shows that the solution presented in this thesis achieves state of the art accuracy rates. However, the ultimate goal with these models is to employ them for practical text coding projects, where the text data might be from different domains or differ in writing styles from the training data. This leads to the first research question to be answered in this chapter:

Research question 1:
How do certain prediction models evaluated with k-fold cross validation perform in real-world application scenarios?

For this project, performance is operationalized as the suitability or fitness of the thesauri generated with the prediction technology for extracting socio-technical networks from different corpora such that the resulting data can be used as input to further network analysis routines. In general, in application contexts, the text data might differ from the training data on many dimensions. In this study, I am testing three of the most common dimensions, namely the time at which some text data were produced, genre, and writing style. Table 84 compares the corpora used in this study, which are introduced in more detail throughout this chapter, to the data used for training the models on the selected dimension. This comparison shows that among the considered corpora, the Sudan data are most similar to the training data, while the Enron email data are most different. Therefore, I hypothesize that the prediction models perform best on the Sudan data, second best on the Funding data, and least well on the Enron data.

Table 84: Comparison of corpora used in application scenarios to used for model training

Dimension	Training Data	Sudan	Funding	Enron
Time	1989	2003-2010*	1984-2006*	2001*
Genre	Newswire	Newswire	Scientific writing*	Emails*
Writing Style	Formal	Formal	Formal	Informal*

* = different from training data

Relation extraction is one among many methods for constructing network data based on text data (for a review of these methods see chapter 3.2.3). However, there is a lack of research on how these different methods compare with respect to their outcome, i.e. the properties of the generated network data. This gap of knowledge motivates my second research question for this chapter:

Research question 2:
How do the network data and respective analysis results obtained by conducting relation extraction including using the entity extractor developed in this thesis compare to alternative methods for constructing network data from the same corpora?

The comparison of network data and analysis result in this chapter is operationalized as follows: based on the experimental results from chapter 2, I had developed recommendations for practical applications of these methods in section 4.1. Based on these recommendations, it seems appropriate to compare networks with respect to their size and the key entities that are identified according to selected network metrics. The latter strategy had also been identified as suitable and was therefore used for comparing networks generated with different coding choices in section 2.7.1. In addition to these strategies for network comparison, the similarity of any pair of network data constructed with different methods is assessed by creating the intersection of these networks in terms of nodes and edges. Since these network data were generated with different methods, which involve different pre-processing steps and pre-processing material such as different thesauri, I hypothesize that the compared network data will not resemble each other. Instead of designing or hoping for convergence of these methods with respect to their outcome, the contribution here rather is to identify the differences and commonalties between the resulting data. This knowledge can help to understand what different views on a network are provided with the tested methods.

In summary, the focus of this chapter is on the impact of methodological choices on network data. This approach is similar to the work presented in chapter 2, where the impact of choices about pre-processing and link formation - all of which also apply to the methods presented in this chapter - was tested. The difference is that while in chapter 2, I used ground truth data to be able to precisely identify these impacts, in this chapter; I use various real world datasets for which no ground truth data is necessarily available. This is possible because in chapter 3, I had used ground truth data to build the prediction models whose performance is contrasted against alternative methods for node identification in this chapter..

5.2 Application Context I: Sudan Corpus

Previous network analytic studies of the Sudan are confined to a few qualitative studies (Elageed, 2008; Lobban, 1975). Conducting participating observations, interviews, or surveys to collect network data about the Sudan and South Sudan is expensive or even infeasible for the following reasons, which might also apply to other geo-political units: the Sudanese population is large (over 45 million people, estimated), the Sudanese people speak over 130 languages, mainly Arabic and/or English (Lewis, 2009), and the literacy rate there is low (61%) (Central_Intelligence_Agency, 2009). As an alternative source of information about this country, one can draw from the large amount of open source text data that are provided about the Sudan.

The presented study of is part of a larger multi-university research initiative (MURI) in cooperation with East Carolina University (ECU) and Rhode Island College (RIC). The goals with this MURI are to (Carley, n.a.):

- Develop theories and computational techniques for modeling the adaptive behavior of groups in asymmetric threat environments.

- Identify and investigate various dimensions of socio-technical networks in the Sudan with a focus on culture.
- Delivering software products that facilitate the fast collection and assessment of these networks.

For the purpose of analyzing socio-technical networks of geopolitical systems, including networks of sub-state and non-state actors, network analysis has been previously employed as a stand-alone method (Erickson, 1981; Hämmerli et al., 2006) as well as a method complementing other techniques, such as regression analysis (Humphreys, 2005). However, direct or remote access to such real-world networks can be hard to impossible for analysts in the cases of covert and past networks, such as illicit groups and bankrupt enterprises (Baker & Faulkner, 1993; Malm, Kinney, & Pollard, 2008). Nevertheless, the networks perspective has been employed to analyze covert organizations and ways of organizing, such as co-offending, trafficking, and white-collar crime (Baker & Faulkner, 1993; Carley, Lee, & Krackhardt, 2001; Howlett, 1980; Reiss, 1988; Sarnecki, 2001; Seibel & Raab, 2003). In these cases, archival data including confidential as well as open source material can help to collect network data (Burt & Lin, 1977). In prior work, people have used text data to answer the following kinds of research questions from a networks perspective:

- Who are the key individuals and groups in a region? (Hämmerli et al., 2006; Schrodt, Gerner, & Yilmaz, 2004; Schrodt, Simpson, & Gerner, 2001)
- How does their importance develop over time? (Carley et al., 2007)
- What dynamics drive the formation of strategic alliances between actors with potentially conflicting interests? (Fitzmaurice, 2000)
- What resources are involved when social agents are in conflict with each other? (Humphreys, 2005)

In order to illustrate the potential utility of coding texts as networks, I provide exemplary, substantive research question such as those outlined above that can be addressed by going through this process and further analyzing the resulting network data. The comprehensive analyses needed to answer these research questions would require separate studies, which are beyond the scope of the thesis. The point with this chapter is to show how the methods and tools studied up to here can be practically employed in a practically useful and efficient fashion.

5.2.1 Data

I put together the Sudan Corpus by using a two-step process that involved downloading documents from the LexisNexis Academic database and deduplicating and cleaning the downloaded files by using software I wrote for this purpose. The same or similar strategies might be useful for others for collecting corpora about countries and geographic regions from open source document collections. These strategies are based on my explorative hands-on work with the data and testing of different choices, such as various search terms and cut-off values. Several heuristics were developed and used as documented herein. These rules might need adjustments when used for building other corpora.

For searching LexisNexis, I used the “power search” as the type of search, “Sudan” as the search term, “major world publications” as the data source, and constrained the search to the “country” category on “Sudan”. A total of 119,859 documents matched these search criteria. As of March 2011, LexisNexis Academic allowed for retrieving 3,000 documents at a time, and downloading 500 at a time; resulting in 246 batches of documents to be manually downloaded. I downloaded the text bodies along with the meta-data that LexisNexis Academic provides. Meta-data are marked by explicit index terms, such as “country”, e.g. Sudan, and “city”, e.g. Khartoum. The meta-data categories and values per category are defined and assigned by LexisNexis Academic without further documentation of this process.

I built a parser to split the batches into individual files and output one text file per article. For each article, the parser identifies the source, publication date, title and actual text body if provided. Since these items are not marked by index terms, I defined data-driven rules for identifying them with high reliability. For cases in which the publication date could not be parsed out, I use the load date, which is a meta-data field, as a proxy. Manually comparing load dates to the publication dates suggested that the load dates are the same or a few days past the publication date.

I set up a database to manage the Sudan corpus. It is common that an article released by one news agency is published by multiple newspapers; leading to redundancy in the reporting of events. I addressed this issue by using the following deduplication strategy: articles with the exact same publication date and title are considered as redundant and were removed. This first round of deduplication reduced the dataset by 4.3% or 5,109 files. The corpus was further reduced to articles relevant with respect to Sudan by keeping only the files that meet both of the following two criteria: (1) The title contains the terms “Sudan*”, “Darfur*”, or “Khartoum*”. The stars represent wildcards. (2) The values for index terms “geography” and/or “country” exceed 90%. These values are determined by LexisNexis and indicate relevance. These two routines together removed another 32,184 or 28.1% articles from the corpus. Further inspecting the data showed that many articles are reports of scores from sports games. I removed articles where the “subject” category contained “soccer”, “basketball”, “tournaments” and “athletes”, which were 1,513 files or 1.8% of the remaining data. Since some articles about sports can be relevant for studying social systems and culture, I kept articles where the “subject” contained “sports”, “Olympics”, “stadiums”, and “arenas”, unless these articles had been removed by the previous steps. At this point, the corpus still had articles that were very highly similar to each other. In order to remove near-duplicates, I disregarded corrections of previously published articles (437 files). Next, I sorted the articles by publication date, title, and source in increasing order. I eliminated those that matched in the first four words of title and were published within a maximum time distance of three days (minus another 1,217 files).

The remaining bodies of the articles still contained index terms and additional information that are not part of the main content and headline, and would be considered noise when performing text analysis. To correct for this issue, I created an instance of the corpus from which I removed the bylines, highlight lines, and copyright notice from each article. Also, I disregarded anything that was not a header or text body, e.g. the phrases “passage omitted” and “Text of report in”. The last step was based on a set of self-defined key words and phrases that indicate the beginning and end of headers and bodies or that serve as indicators for irrelevant lines and phrases that are intermitted within the body.

Next, I added a sentence mark at end of each headline. For the vast majority of articles, this helps to make the headline look like a real sentence to any subsequently used text mining routine. However, if the headline already has a sentence marker, e.g. a question mark, this operation resulted in two delimiters for an end of a sentence.

Finally, I checked if the cleaning techniques had reduced any articles to something not useful for text analysis anymore, such as nothing but section markers or image captions. Going from the smallest to the largest texts, this step eliminated 12 more articles. In total, the cleaning techniques reduced the corpus by 33.8% or 40,471 articles to 79,388 files. Table 85 shows the number of articles per calendar year in the final Sudan corpus.

Table 85: Articles per year in Sudan corpus

Calendar year	2003	2004	2005	2006	2007	2008	2009	2010
Number of articles in corpus	4,507	10,059	7,837	11,076	12,243	10,713	10,410	12,543

5.2.2 Network Data Construction Methods

The same network data construction methods were used for the three different application scenarios/ corpora in this chapter is possible. For the Sudan corpus, the following four methods were used:

1. Perform text coding with the data to model process (D2M) in AutoMap (explained in section 5.2.2.1). This process involves the construction of a thesaurus.
2. Same as above, but with the difference of using a thesaurus generated by the entity extractor built in chapter 3 (5.2.2.2).
3. Construct network data from meta-data contained in the Sudan corpus (section 5.2.2.3).
4. Work with subject matter experts to constructed network data that can be considered as ground truth data (section 5.2.2.4).

5.2.2.1 Network Data Extraction from Texts Using the Data to Model Process

The data to model (D2M) process was defined by Carley et al. (2011), and is designed for going from texts to multi-mode, socio-technical networks to analysis results. The process is still evolving and has been used for multiple text coding projects in CASOS. Also, the process has been tied to the CASOS tools, namely AutoMap (Carley, Columbus, et al., 2011) and ORA (Carley, Reminga, et al., 2011). These tools are publicly available and are also described herein as needed. Next, I explain the D2M process at its current state and how it is used in this chapter. The D2M process starts with text data collection:

1. Collect a text corpus (described in section 5.2.1).
2. Clean the text corpus (described in section 5.2.1).

The next set of steps of the D2M process is designed for extracting relational data from texts. These steps involve various pre-processing routines, which are further explained in the next section, and are provided in AutoMap:

3. Create thesauri and/ or adapt existing standard and domain thesauri such that they are appropriate for the given research question, domain and dataset.
4. Review and revise the thesauri.
5. Extract meta-networks from the corpus.
6. Review the network data and based on that, revise the thesauri.
7. Recreate meta-networks from the corpus.
8. Iterate through steps 4 to 7 until the network data seem appropriate.

Once these steps are completed, the extracted data are post-processed in ORA to add geo-spatial information to the extracted networks (step 9). Next, network analysis is performed on the data (10). Then, analysts can use the results to suggest potential interventions (11). Finally, simulations are run on the data to explore what-of scenarios and potential interventions (12). For the application scenarios presented in this chapter, I perform steps 1-8 and 10 as they are relevant for the purpose of this chapter.

5.2.2.1.1 Thesauri: Background, Usage and Construction

The key resource needed for extracting meta-networks with the D2M process are thesauri. A thesaurus, in its simplest form, is a table with two columns that associates text-level terms (first column) with concepts (second column). When applying a thesaurus, the text data are searched for the terms listed in the thesaurus, and any matches are replaced with the respective concepts. In order to build thesauri, a combination of data-driven NLP techniques, external resources such as gazetteers, and previously generated thesauri is typically employed. In AutoMap, the NLP techniques available for this purpose include the identification of terms (unigrams and bigrams) with high absolute and weighted frequencies (Diesner & Carley, 2004) and the automated detection and classification of nodes (Diesner & Carley, 2008a). Some of these techniques are computer supported in AutoMap, i.e. they require manual steps, while others are fully automated. To give an for computer supported routines, before the prediction models presented in chapter 3 were added to AutoMap, the process for detecting multi-word units involved generating a bigram list, which contains all adjacent pairs of words and their cumulative frequencies. The disadvantages with this approach were that the output had to be screened by a person for meaningful two-word units and the detection of longer units was not supported. Note that in general, alternative fully automated methods and tools for N-gram detection are available.

A thesaurus can be used to normalize data as shown in the examples in the next paragraph, or as a positive list or filter, which means that all text terms not occurring in the thesaurus are dropped from the text data. More specifically, in text coding, a thesaurus serves four main purposes, which may overlap:

First, it converts explicit literal mentions of concepts into those concepts, e.g. “cocoa beans” into “agricultural_crops”. Used in this way, a thesaurus represents a taxonomy, which classifies terms into concepts. Second, a thesaurus supports coreference resolution by mapping different spellings, variations, and synonyms of a concept to one consistent key identifier for this concept. For example, “Al-Bashir”, “Omar el Bashir”, and “Omer Hassan Ahmed al-Bashir” can all be mapped to “Omar_al_Bashir”. Third, a thesaurus helps to disambiguate terms. This works for terms where capitalization signals a difference in meaning (capitonyms), e.g. “rice” (crop versus person with that last name). Disambiguation via a thesaurus can also be achieved for terms that have the same spelling but a different meaning, i.e. homographs, which include homonyms, heteronyms, and polysemes. However, disambiguating homographs via thesauri is only feasible if and only if the embedding of the term into the context of a short phrase is sufficient for differentiating their meaning, e.g. “upper arm” versus “arm dealer”. Forth, a thesaurus can be used to convert n-grams into unigrams. This is typically done by replacing the spaces between the constituents of an n-gram with an underscore as shown in the examples in this paragraph.

Thesauri that are more advanced than the basic two-column data structure contain additional columns that specify the type and further subtypes and attributes of entities. I herein refer to these additional pieces of information about an entity as “categories”. For instance, “Omar_al_Bashir” might be categorized as an entity of the type “agent” with the specificity value “specific” and the subtype “political”. Thesauri that associate terms with categories allow for text coding and subsequent analysis on multiple levels of aggregation, and also for fine-grained analysis and filtering.

Traditionally, thesauri have been created by reading through some (Glaser & Strauss, 1967) or all (Gerner et al., 1994) of the text data to be analyzed in order to identify the terms relevant for a given project, and then associating these terms with concepts and categories. Sometimes, the relevant concepts are predefined, e.g. if they are derived from theory or when ontologies or taxonomies are used. Various computational solutions exist for assisting the user in this task; many of which have been developed for qualitative text coding according to the grounded theory methodology (Lewins & Silver, 2007) and for event coding in the political sciences (Gerner et al., 1994).

Thesauri are typically created through an iterative process of testing and modification. Sometimes, external resources can be used to build or extend thesauri. For instance, Appendix A of the CIA World Factbook lists acronyms that are commonly used for various organizations, such as “WHO” for “World Health Organization” (Central_Intelligence_Agency, 2009).

There are two main advantages with thesauri: first, they allow for working with a controlled vocabulary. Second, they support the consideration of subject matter expertise for text coding. This means that while experts are able to define terms that represent relevant concepts in a domain and also to categorize terms, these concepts and categorizations might not be identifiable with statistical NLP techniques.

Thesauri involve several limitations: first, they can be outdated, incomplete, insufficiently discriminating between the different meanings of terms, and miss the typos occurring in real data. The deterministic nature of a thesaurus can be improved by not only searching for hard matches, but also soft matches in spelling via string similarity algorithms (Cohen, Ravikumar, & Fienberg, 2003). Second, since thesauri are typically built for specific domains, genres or datasets, they can be expected to perform less accurately on new corpora. Finally, building thesauri built by hand or in a computer assisted fashion is very costly in terms of effort and time.

5.2.2.1.2 Construction of the Sudan Master Thesaurus

For this study, I am using a thesaurus herein referred to as the Sudan “master thesaurus”. This thesaurus was built by various members of CASOS over multiple years by integrating multiple thesauri previously built at CASOS and elsewhere, enhancing the resulting file with the D2M process, and repeatedly cleaning and enhancing the file. These steps were mainly conducted by individuals other than me inside and outside of CASOS, and no complete documentation exists for this process. Therefore, I consider the master thesaurus as a given input. This situation might also apply to other real-world text analysis projects where dictionaries from external sources are used.

This section describes how I refined and enhanced the Sudan master thesaurus. Out of the different thesauri that I built for this chapter, the Sudan master thesaurus required the most amount of effort in terms of time and labor for cleaning and manual validation. The resulting file can serve as a starting point for building thesauri that can be used for analyzing data on other geo-political entities and other news wire corpora, which is a main application domain for thesauri in CASOS. For these two reasons, I use the master thesaurus not only for this application scenario, but also as a look-up dictionary for constructing the prediction models in section 3.3.2.4.

I want to mention two particularly important thesauri that had been previously integrated into the master thesaurus: first, the counter-terrorism agent thesaurus (CT agent thesaurus), which is a collection of entities of the type “agent” that are relevant in the context of counter terrorism. This file has been constructed and verified by subject matter experts (Gerdes, 2008) and accounts for 20.6% of all agent entries in the master thesaurus. Second, the rapid ethnographic retrieval (RER) thesaurus, which was was built by our project partners at East Carolina University. This file associates terms with concepts that subject matter experts have identified as being crucial for answering questions about the culture of groups and societies. These term associations result from both, theory and empirical work in anthropology and sociology (Carley, Lanham, et al., 2011). Many of the RER terms are based on the “Human Relations Area Files” (HRAF), which are a classification schema for information about human behavior and culture. The HRAF are widely used for anthropological research. The RER thesaurus ranges across multiple entity classes, and provides 2.7% of the entries in the master thesaurus.

All terms and concepts in the master thesaurus, except for a list of about 13,000 universities, are in lower caps. This eliminates the need to enter terms twice if they can occur either way, but at the same time disables the possibility of word sense disambiguation of capitonyms.

The raw version of the master thesaurus that I use is from May 25th, 2011. Towards the end of the cleaning and refinement process described in the following I was given an updated RER thesaurus with entries for the task, resource and knowledge class as well as a list of about 13,000 universities that are classified as organizations with the subtype "educational". Integrating these files with the master thesaurus required repeating all cleaning steps described in this chapter for these two files, and deduplication all impacted entities classes again. The numbers presented in this chapter are adjusted for these additional steps. This limitation to efficient scientific work reflects the nature of practical text coding applications: thesauri are ever evolving instruments that need to be adjusted for time, domains, and writing styles, among other criteria.

The master thesaurus has seven columns: the "terms" (229,998 lines), one "concept" per term, the "meta-network category" that the concept maps to (for 99.4% of the concepts), a "subtype" per concept (for 14.7% of the concept), and the "city", "state" and "country" for the entries from the university file if available. Table 87 shows the distribution of terms across categories. I cleaned and enhanced this file as follows:

First, I used a CASOS tool that helps to remove lines that contain illegible characters in the term and concept column. This tool converts characters from the UTF encoding set to the respective ASCII characters while leaving all other ASCII characters untouched. Terms removed included "x x•x"x¤x•x§" and "D±N€NƒD½DµN". Those entries resulted from scraping webpages and moving files between different encoding sets without adjusting for the character set. This step reduced the number of lines by 19.5%. Of those lines removed, 97.6% were from the "location" class, and another 1.6% from the "agent" class.

Next, I manually fixed all typos in the meta-network categories (N=107, where N is number of lines). This is important because otherwise these classes would be considered as additional categories. Also, I removed all entries marked as "ignore" (N=18), which were leftovers from a prior (to this thesis) round of editing.

Then, I checked all entries that had an underscore between words in the term column (N=2,751), which are the result of previous issues with merging and deduplicating thesauri. Underscores are only supposed to occur in the concept column and are there to covert n-grams into unigrams. Of those entries, I removed all but those from the RER thesaurus and fixed the RER entries (171 entries with underscores kept).

At this point, the thesaurus still had several entries that were noise and featured certain symbols. Again, those entries might result from collecting data online and from moving information between different character encoding sets, among other reasons. I manually worked through these entries:

Question marks (N=569): I vetted 14 of the entries as relevant and unproblematic; most of which were speech acts and abbreviations used in web talk, such as “wuf?” (an abbreviation for “where are you from?”). I fixed another 38 entries by removing the question marks and removed the remainder as it was noise.

Quotation marks (N=480): I kept 48 of those entries; some of which needed some manual fixing. The rest was dropped because they were noise. The maintained entries are from the “agent” class, such as “haji neamatullah "shirdai" khan”, and terms representing universities, such as University "”Dzemal Bijedic" of Mostar”.

Digits: When the D2M process is used to retrieve potentially relevant entities from text data, entries with digits are removed as those entities are often considered as noise. Since we had no data on how appropriate this strategy was, I went through all entries in the thesaurus that contain a digit within the term (N=3,012). Of those, 49.5% are industry codes, e.g. “naics111140 wheat farming”, and news ticker IDs, e.g. “9501 (tse)”; both of which I did not attend to. Out of the 1,527 remaining entries, I vetted 39.7% as noise and dropped them, 32.6% as relevant and correctly formatted, and 27.7% as relevant yet problematic. I fixed the problematic cases, e.g. by removing the digit from the term or correcting the meta-network category. All entries that I did attend to were added back into the master thesaurus. Table 87 shows that digits are a meaningful constituent of more than 50% of the entries that comprise digits (excluding industry codes and ticker IDs) such that dropping them entirely would cause a loss of information.

In total, the handling of the entries that contain certain symbols shows that 90% or more of the terms comprising question marks and quotation are noise, while digits are a relevant component of about every other impacted entity.

Table 86: Overview on entries with digit(s) in term values (excluding industry codes, ticker IDs)

Meta network category	Number of entries with digit(s) in term	Number of entries with digit(s) after digit cleaning	After cleaning, entries with digit(s) being relevant	After cleaning, entries with digit(s) being irrelevant
Agent	263	151	22%	78%
Atribute	7	0	0%	0%
Event	89	58	84%	16%
Knowledge	151	86	59%	41%
Location	290	188	62%	38%
Organization	534	307	60%	40%
Resource	148	89	61%	39%
Task	35	42	29%	71%
Blank	10	0	0%	0%
Total	1,527	921	54%	46%

After the symbol handling was finalized, I manually defined concepts and meta-network categories for all uncategorized terms (N=1,024). There is no explicit code book that would guide this process, but several guidelines (Carley, Columbus, et al., 2011) and plenty of norms have been established in the CASOS center for this process. I adhered to these norms, built upon my experience with plenty of previous text coding project in CASOS, and double checked on the cases that I was uncertain about with the director of CASOS, Dr. Kathleen M. Carley.

Next, I worked through the entries in each entity classes individually. Doing that for the agent class took the most effort, and the steps required there do not necessarily generalize to the handling of the other entity classes. Therefore, I describe this process separately, followed by a general description of problems and solutions for the other nine entity classes.

5.2.2.1.2.1 Agents

Most of the problems with the agent entries were cases in which instances of “roles” were lumped together with reference to specific agents, such as “president omar al-beshir”. Also, for all agents, we want to be able to distinguish between specific (omar al-beshir) versus generic (president) instances. However, of the 29,690 agent entries, only 1,789 (6%) were marked as “specific”, and 30 as “generic”[14]. Moreover, instances of “roles” and “generic agents” mainly overlapped. Another minor issue with the agent class was that several concepts contained spaces, which I replaced with underscores.

In order to split up entries composed of generic and specific references to agents and to classify all entries into either one specificity type, I started by manually reviewing the existing CASOS roles file. This file has 741 entries. I decided to remove 18 of them, mainly because they often occured as part of proper noun phrases, i.e. specific agents, e.g. “khalif”. I built a tool that applies the roles file to the terms and concepts columns of a thesaurus and separates roles from specific agent representations per line and column. Next, I went through all agent entries and took everything that did not represent a specific agent out into a separate file (delete list). This delete list contained 2,820 entries, some of which were additional roles, while others were noise.

Several types of conflicting cases were less straightforward to handle: some instances of roles are often part of proper names, e.g. “pope” (“pope john paul”), “father” and “prophet” in a religious context, or “khalif” and “khalifa”, e.g. Ayad Futayyih Khalifa al Rawi. Removing the role from the name would not allow for mapping this name anymore to the text data, but might still be helpful for cleaning up other names. Also, some roles overlapped with common proper names, such as “king” in “martin luther king”, where removing “king” would also alter the proper name

[14] Two more agent entries had the subtypes “corporate” and one as “non-corporate”.

in an undesired way. Furthermore, some roles coincide with common nouns and noun phrases, such as “west” in “allen west”, where mapping every instance of “west” in text data to this particular agent might more often be wrong than right. For these scenarios, I made case by case the domain of decisions based on which usage of a term (role or any other) seemed more common for news wire data. Applying the resulting extended delete list to the agent entries did impact 34.9% of the terms, 12.8% of the concepts, and 35.4% of all agent entries. Out of all term-concept pairs that were subject to this process, 6.5% were reduced to empty pairs. It is noteworthy that only 8.6% of the entries from the CT agent thesaurus were impacted by the role removal process, which indicates that these entries had already been subject to cleaning procedures and consistency checks.

In AutoMap, once a thesaurus has been constructed or changed, co-reference resolution has to be performed on the thesauri in a manual fashion. This involves mapping synonyms to unique node names. Also, since AutoMap does not yet disambiguate terms based on capitalization or part of speech, a person has to decide which meaning of a capitonyms and homographs to assign to all instances of these words, e.g. whether to code “rice” as a person in the sense of the politician or a resource in the sense of food (AutoMap does not distinguish thesaurus terms based on capitalization). The master thesaurus supports co-reference resolution by associating different variations of a name with a unique spelling of that name. Several pseudonyms, aliases and noms de guerre are also handled by the thesaurus. Since the cleaning routine described above had impacted the terms and concepts, I had to redo the co-reference resolution. In fact, both, the CT agent file as well as the other agent entries contained cases where one term was mapped to multiple concepts in the original master thesaurus. I iteratively developed and implemented a rule-based approach to solve this problem:

- All comparisons are performed on the level of exactly matching letters and numbers, but not symbols.
- For all cases in which multiple occurrences of one term map to more than one concept, the concept from the CT agent file is used if the term occurs in the CT agent file, otherwise the most frequent concept is used.
- In the case of a tie, the term that first occurs in the alphabet is used.
- For unigrams, additional rules are applied: conflicts for unigrams occur if one part of a name, e.g. Smith, is mapped to multiple combinations of a first name and a last name, e.g. Amy Smith, Betty Smith, Cary Smith, etc.. For first names, it is hard to tell which full name it is to be associated with. Therefore, unigrams are associated with the concept from the CT agent file if the unigram occurs only once. Otherwise, the unigram is translated into itself.

Next, I deduplicated all agent entries by removing those entries that were identical in terms and concepts. The deletion and co-reference resolution process had caused several terms to shorten in the number of tokens, which implies the risk of mapping a meaningless or overly common term to an agent, such as “john” (unclear which “john” is meant). I reviewed all terms and concepts that had a length of four characters or less (N=686), and removed 27 of them as they were noise. During this process, I found ten more terms that had been reduced to roles. I removed those lines, but did not add the roles to the role file since those term represented some of the difficult cases described earlier.

Next, I manually classified all entries in the roles file as the respective meta-network category unless they were noise. Most of them were assigned to agent of subtype generic or to attributes.

Finally, I checked the agents file against a list of tribes in Sudan and removed one matching entry from the agent file (“subayh”). This would have been a false positive in the agent class.

5.2.2.1.3 Using the Master Thesaurus for Extracting Meta-Networks

Once the Sudan master thesaurus was built, I used it as part of the D2M text coding process in AutoMap. Since the corpus and thesaurus are sizable, I used the script version of AutoMap for processing. With this version, the user fills out a script that specifies the coding choices and input and output directories. In order to choose appropriate coding choices for this project, I drew from the knowledge gained in chapter 2 and from consultations with other members in our group who were also processing the Sudan corpus and other text datasets about large-scale, geo-political entities. I specified the following coding choices:

- *Cleaning of all texts*: this routine deduplicates texts, removes meta-data, corrects typos by applying a thesaurus of common typos, and expands contractions and abbreviations by using thesauri.
- *Thesaurus application*: the master thesaurus described in the previous section was applied such that only entries matching the thesaurus are kept in the data (thesaurus content only option) while maintaining the original distances between concepts (rhetorical adjacency option). Comparisons between text terms and thesaurus entries are performed on a lower case basis. All concepts in the output data are also in lower case.
- *Meta-network extraction*: AutoMap uses the windowing technique for link formation. The parameters taken into account for window-size specification include the text unit, such as sentence or paragraph, and the number of words. Based on the experimental results and respective practical implications for appropriate window sizes from chapters 2 and 4, I used a window size of seven. Also, I allowed for the windows to span across a

sentence. In order to address the potential risk of finding false positives, I coded roles and attributes not as instances of node classes, but as attributes of nodes.

The output from this process are directed, weighted graphs that are output in the DyNetML format (Carley, Reminga, et al., 2011); a XML format developed for describing graphs. One DyNetML file is output per input text file. I consolidated these outputs as follows: all files that were published in the same calendar year were aggregated into one DyNetML file per year. This requires each filename to contain the time stamp from the article in a specific format (yyyymmdd). I used the publication data of articles as the timestamp. A limitation with this approach is that the actual event may have happened prior to the publication data. Each resulting DyNetML file represents all the nodes and edges that were found in all of the text files per year. If a node or edge were found more than once, their initial weight of one was increased accordingly. Once this process was completed, the DyNetML files were loaded into ORA.

Inspecting the network data files in ORA showed that many nodes still appeared as multiple mentions, i.e. they represented the same entity, but had different node IDs and thus occured as multiple nodes. For instance, there were still 18 different nodes that all represented Omar al-Bashir. I used the following strategy for conducting another round of co-reference resolution, now on the node level: first, I loaded and applied attribute files that assign a specificity value to nodes where available. I had built these attribute thesauri as part of the master thesaurus and also for my previous work on coding the Sudan data. Except for the agent class, these thesauri did not cover all nodes in the networks. Therefore, I labeled all nodes from the organization class that had a frequency of 1,000 and more in the union of all annual networks with the best fitting specificity value. The number of 1,000 was chosen as an artificial cut-off point. Ideally, one would want to assign a specificity value to all entities, but since this process has to be done manually, such procedure would not be feasible for a single person in a reasonable amount of time. Next, I selected all agents and organizations with the specificity value “specific”, and for each of these nodes with a total occurrence of 1,000 times and more, I checked if they can be merged with any other node from the same class and of any frequency, including frequencies of less than 1,000. The resulting node merging lists can be stored, but need to be applied to every network and node class individually in ORA. In total, just the process of assigning specificity values and conducting co-reference resolution on the node level took about four days.

In summary, in comparison to the original agent portion of the master thesaurus, the reworked portion contained 19.5% less unique agents and term-concept pairs (N=23,832), and 5.0% less unique concepts (N=19,387). All remaining unique agents are specific ones - an increase by 22,043. Preparing the agent entries of the master thesaurus involved several limitations:

First, terms that represent generic as well as specific agents were not removed from the file in order to not to lose this information altogether. An example would be “christian”, which can be a first name or a person that adheres to the Christian religion.

Second, translating unigrams into themselves causes a loss of precision in some cases, while in others, it avoids the mapping common first names (paul, bill, mark) or common other words (ban, rice) to one specific agent.

Third, terms that only differ in symbols are not considered as being identical, such as “hassan yemen al-rabiai” versus “hassan yemen al rabiai”. I chose this rule because differences in symbols often signal different agents, or this strategy would merge a term with a non-agent term, such as “sa-id” and “sa’id”; both of which are common first names in the given domain.

Forth, the co-reference resolution approach is not optimal and also incomplete. On average, each agent concept in the final master thesaurus maps to 1.2 terms. For example, “omar hassan al-bashir” is mapped to “omar_al_bashir”, while “omar hassan ahmad al-bashir” is mapped to “omar_hassan_ahmad_al_bashir”, even though many variations of this name are collected together under the latter and more common spelling. The rule based consolidation approached used herein can only partially alleviate those issues. Moreover, in many cases, it is not obvious if two similar names really represent the same person. Further resolving this limitation would require subject matter expertise and further manual work.

While the first three limitations are classic caveats of rule based systems, the forth one is a known shortcoming of thesauri. Furthermore, the first two limitations are specific to the agent entries, while the last two ones also apply to cleaning other entity classes, which is described next.

5.2.2.1.4 Limitations of Working with Thesauri

In general, the manual and semi-automated verification and correction of a thesaurus as demonstrated in this section serves the validation of a thesaurus and the improvement of the quality of the thesaurus. However, working with thesauri involves several limitations, which are described in the remainder of this section. These issues are mainly due to the fact the master thesaurus was built, maintained and extended over years based on multiple sources and rules and by multiple people and teams from multiple organizations, which is a realistic and common scenario. Working through the remaining nine entity classes (organization, location, resource, knowledge, task, event, time, belief, attribute) revealed several common issues. These issues are mainly due to the following issues, which may overlap:

- Homonymy of terms and concepts.

- Gathering of data from external sources, such as the web (potentially messy) and structured databases (cleaner than the web).
- Integration of information from various research groups, such as the cultural indicators (RER) from ECU with the CASOS thesauri.
- Pre-processing of the text data prior to thesaurus construction.

In the following paragraphs, I describe some of these problems in more detail: first, concepts considered in the thesaurus are sometimes represented by very common terms ("conflict" by "against") or by terms that have another yet more common meaning ("well" coded as "water"). These two problems were solved by removing overly common terms from the thesaurus, such as "go", "take", "will" (intended sense was a declared intention) and "me" (personal pronoun and abbreviation for the state of Maine).

The second issue results from AutoMap coding every distinct term into only one concept; with ties being broken alphabetically. This is problematic for terms that map to more than one distinct relevant concept, such as "fur" to one of the main tribes in Sudan as well as to a natural resource. The same problem applies to acronyms and abbreviations which represent multiple entities. In these cases, I chose the anticipated and more frequent meaning in the context of the text corpora used herein.

Third, various concepts appeared in multiple meta-network categories, such as the "oslo accords", which is short for the "Declaration of Principles on Interim Self-Government Arrangements", and was coded as knowledge (in the sense of a document) as well as event (in the sense of the meeting). For these cases, I developed data-driven rules that I adhered to. In total, the master thesaurus contained over 1,000 conflicting cases where the same term was assigned to more than one concept, or the same concept assigned to more than one categories. Resolving these issues required working through them on a case by case basis by hand. For some homonymous terms where the different meanings (concepts) were each highly relevant, the less common meaning was eliminated. For instance, I dropped "turkey" coded as "livestock" in order to keep "turkey" coded as a location. Some term to concept assignments were kept since they occurred frequently with the intended meaning in the corpora used herein, e.g. "general" as a military rank, but these assignments might not be appropriate for other datasets. Furthermore, decisions on several terms required substantial subject matter expertise. For example, there were several hundred terms coded as a person and an organization (e.g. "wazir"), or as a person and a location (e.g. "bahr el ghazal"). For these cases, the most appropriate assignment was not obvious to me. Resolving these issues required substantial additional research.

Fourth, many terms that were picked up by automatic entity extraction techniques when building the thesaurus contained irrelevant words in addition to the relevant ones, such as verbs as well as the names of months and days of the week as part of noun phrases. I removed those when I found them and where it seemed appropriate.

Fifth, several sections of the master thesaurus were retrieved from external webpages. In general, extracting relational data from the web has become a useful and popular strategy for filling relational databases (Cafarella et al., 2006). However, scraping the web for collections of terms and concepts can result in the retrieval of large numbers of additions to the thesaurus, but these entries include noise that requires further inspection and cleaning. For example, many of the locations were collected from resources that include the foreign translation of location names, which sometimes coincide with common English terms.

Sixth, the creators of different thesauri had not always used the same guidelines or codebooks for associating terms with concepts. For instance, the RER thesaurus often codes roles as resources, such as “laborer”, while the CASOS role file considers them as roles. Also, the RER thesaurus considers diseases as knowledge, which would be appropriate in the context of research papers, while the CASOS thesauri consider them as a resource in the sense of something that one can acquire. Since the RER thesaurus was built by experts it was given precedence in most cases. Many of these conflicts have no right or wrong solution to them. The choices made are based on norms and guidelines specific to an organization or a field and on the context of the corpus to which the thesaurus is to be applied.

Seventh, the master thesaurus includes stemmed versions of terms. The problem here is that some morphemes coincide with other common English terms. This issue particularly applied to location names that were retrieved from external digital resources. Also, the stemmers that were used are designed for English text data (Diesner & Carley, 2004), such that errors on applying them to foreign words can be expected.

After reviewing the entries per entity class and correcting for the outlined issues where possible, the revised master thesaurus required performing disambiguation and deduplication again such that some of the issues outlined above had to be addressed again. I also kept one thesaurus per entity class since those contain more entries than the consolidated master thesaurus. In order to test the quality of the revised master thesaurus and to check for further noise terms and inappropriate pairings, I applied the thesaurus to the Sudan corpus as follows: I generated a term distribution list that specifies the cumulative, observed frequency of each term and concept, and how many texts they occur in. I inspected all occurrences with a frequency of 1,000 and higher (N=1,607), and fixed all problematic entries. Repeating this process one more time and inspecting the thesaurus again suggested that the quality of the thesaurus was sufficiently high at this point.

Overall, the thesaurus cleaning procedures had major impacts on the master thesaurus as summarized below. Table 87 further provides a quantitative overview on these impacts.

- The number of entries in the master thesaurus was reduced by over 26%. While some classes are reduced by even larger ratios, the role class and to a lesser degree also the attribute class were extended.
- Over 43% of the entries in the master thesaurus were changed in one or more column. This means that the qualitative effect of cleaning the thesaurus is larger than the quantitative impact.
- More than 76% of the entries in the revised file were taken from the original file with no changes, but this ratio differs widely depending on the entity class: in fact, for six out of the ten classes, more than 85% of the entries in the revised file are from the original file. This means that while large numbers of entries were dropped from each original class, the remaining original entries make up the bulk of the entries in the revised class. However, for the classes of agent, attribute and role, almost all entries have gotten changed or added after dropping noisy and erroneous entries.

Table 87: Size and categories of master thesaurus, original and revised

Meta network category	Number of entries in master original	Number of entries in master revised	Change in number of lines from original to revised	Number of lines identical between original and revised	Entries in revised retained unchanged from original	
					base: original	base: revised
Agent	30,822	24,160	-22%	995	3%	4%
Attribute	669	768	15%	0	0%	0%
Belief	268	271	1%	260	97%	96%
Event	1,898	1,665	-12%	1,633	86%	98%
Knowledge	5,741	4,621	-20%	4,142	72%	90%
Location	147,885	101,163	-32%	100,458	68%	99%
Organization	32,232	29,199	-9%	17,240	53%	59%
Resource	5,631	2,345	-58%	2,005	36%	86%
Role*	73	1,946	2566%	42	58%	2%
Task	3,647	3,653	0%	3,267	90%	89%
blank	1,024	0	-100%	0	0%	0%
wrong categories	108	0	-100%	0	0%	0%
Total	229,998	169,791	-26%	130,001	57%	77%

* in revised: agent generic

Two more limitations apply to the thesaurus revision process: first, all cleaning and rule creation described herein was done by a single person (me) in consultation with the people involved in handling our thesauri and my advisor. Any errors that I did not spot remain in the data until somebody else finds them. Second, building, refining and extending thesauri is very costly in terms of time and human efforts: working through 500 lines took me about one hour on average for most of the processes described here. Altogether, revising the master thesaurus took about six work weeks. Adjusting the master thesaurus to another dataset or domain, or building an entirely new thesaurus, is likely to involve significant time costs of several days, weeks or months. However, once this work is done, using the thesaurus is efficient: the total time costs for coding texts as networks in AutoMap and consolidating the files as described in this section were about a day and a half. Using the revised master thesaurus as is will not increase time costs beyond the processing needed for AutoMap. Moreover, in AutoMap, a plethora of previously generated thesauri are provided to end users. Those are general thesauri that handle the conversion from British to American English, expansion of contractions and common abbreviations.

5.2.2.2 Network Data Extraction from Texts Using the Data to Model Process and the Entity Extractor

The same process as outlined above was repeated for generating network data with the D2M process, but with one change to it: I replaced the Sudan master thesaurus (mixture of manual and computer-supported construction) with a thesaurus generated by applying the entity extractor developed in chapter 3 to the Sudan corpus (fully automated initial thesaurus construction). I refer to thesauri generated this way as auto-generated thesauri. Inspecting the auto-generated thesaurus for the Sudan data and a first batch of network data generated by using this thesaurus suggested that the auto-generated thesaurus cannot be used as is to retrieve quality network data, but also needs further cleaning. However, the auto-generated thesaurus featured different issues than the master thesaurus such that different strategies seem to be necessary for handling them. Thus, I refined the auto-generated thesaurus as describe below. This description might also serve others who use the entity extractor in AutoMap to convert thesauri suggested by the entity extraction technology into quality text coding tools.

Refining the auto-generated thesaurus was an iterative process: I implemented a particular change, used the modified thesaurus to generate network data using the same process as described in section 5.2.2.1.3, inspected the thesaurus and the network data[15], made further

[15] Since the thesaurus format in AutoMap accepts one attribute per entity, I stored the additional attributes (subtype, part of speech value) as separate files and added them into the DyNetML files in ORA.

changes to the auto-generated thesaurus, and repeated this process. The steps described in this section are not all of the changes I tested, but those that I assessed as being effective and leading to the intended improvements without causing unintended side effects. Also, I tried different orders of applying these steps. The sequence of routines described in this section is the ordering that led to the best quality of the auto-generated thesaurus.

For thesaurus generation, I used class model 4, which outputs a class label, specificity value, and subtype value for each identified entity (for details on the class models see Table 60). The output from this routine further contains the part of speech for each constituent of an entity and the frequency with which each entity (case-sensitive, agreeing class label, specificity value, subtype, and part of speech) has been identified in the text data. The auto-generated thesaurus had 502,485 unique, regular entries with a cumulative frequency of 5,380,091 instances, and another 28,922 additional suggestions (for details on the additional suggestions see section 4.2). Since the number of regular entries was already large, many of the additional suggestions were already contained in some form among the regular entries, and many of the additional suggestions seemed only tangentially relevant, I decided to remove them from the auto-generated thesaurus.

In order to assess the quality of the auto-generated thesaurus in a practical application setting, I manually reviewed the suggested entries per category (total of 44 categories). Table 88 lists these categories along with their accuracy obtained during k-fold cross-validation. These accuracy rates serve as a point of comparison here (for details on formal model evaluation see section 3.4.7). The table also contains the cumulative sum of retrieved instances per class and my assessment of the prediction accuracy per class in the application context, i.e. the Sudan corpus. I performed this assessment in a qualitative way: I screened the entries per class; especially those with high frequencies, and categorized each class as having good, medium or bad prediction accuracy in the application domain. Ultimately, such an evaluation should be performed by multiple people to avoid intra-coder reliability issues and biases. However, the presented initial evaluation serves two purposes: first, to identify general issues with the auto-generated thesaurus and to understand how these issues relate to problems identified for the master thesaurus (described in section 5.2.2.1). Second, to understand which issues are corpus specific and which generalize across various application scenarios. This second question about the generalizability of positive and negative aspects of the auto-generated thesauri requires the validation of such thesauri in various application domains, which will be demonstrated for three different datasets in this chapter.

Table 88: Application of prediction model to auto-generate thesaurus for Sudan corpus

Class labels	**K-fold cross validation**	**Application to Sudan data**	
Meta-network category, specificity, subtype	Accuracy	Size: Number of examples in thesaurus	Assessment of quality
resource, na, money	97.7%	28,757	good
location, specific, country	97.0%	606,204	good
org-att, specific, nationality	93.8%	145,578	good
attribute, na, numerical	93.4%	394,769	good
time, na, na	93.4%	396,072	good
event, specific, war	92.6%	2,280	good
agent, specific, na	92.3%	200,658	bad
organization, specific, gov.	90.8%	136,919	good
org-att, specific, political	90.5%	807	good
agent, generic, na	90.2%	882,345	good
organization, generic, corp.	88.7%	283,014	good
location, specific, city	88.1%	157,603	good
organization, specific, corp.	87.2%	854,630	medium
location, generic, country	87.1%	126,048	good
location, specific, state-prov.	85.4%	7,059	good
organization, generic, gov.	81.4%	71,840	good
organization, specific, edu.	77.8%	15,645	good
location, generic, city	77.7%	24,098	good
knowledge, specific, law	77.5%	48,340	good
organization, generic, edu.	72.7%	5,826	good
location, specific, other	71.8%	34,687	good
resource, generic, product	71.7%	96,935	good
event, specific, na	69.0%	9,917	medium
location, generic, facility	67.9%	60,165	good
organization, specific, other	67.1%	155,225	good
attribute, na, age	66.9%	37,860	good
organization, specific, political	63.8%	15,408	good
resource, na, substance	62.0%	36,810	good
organization, generic, other	61.6%	67,556	good
org-att, specific, religious	59.6%	2,517	good
location, generic, state-prov.	52.9%	34,354	good
resource, na, disease	50.8%	9,944	medium
knowledge, specific, language	50.0%	3,484	good
location, specific, facility	49.8%	35,929	medium
knowledge, specific, art	48.5%	312,947	bad
organization, specific, religious	48.5%	15,896	good
resource, na, plant	48.5%	2,918	good
organization, generic, political	48.3%	469	good
organization, generic, religious	47.1%	4,238	good
resource, na, animal	40.4%	8,598	good

org-att, specific, other	34.4%	15,621	good
task, na, game	29.6%	378	good
resource, specific, product	28.0%	26,968	bad
location, generic, other	18.8%	2,775	good

During this assessment, I made the following observations:

First, overall, many of the identified entities and associated classifications seemed relevant and correctly labeled, respectively.

Second, some categories were particularly error-prone. Most of those errors were cases in which relevant entities were picked up, but assigned to the wrong category. Especially agents with the specificity value "specific" were particularly likely to show up in other categories, mainly as specific knowledge of subtype art and as specific organizations. The latter issue was also observed with the master thesaurus, where deciding on the right category for some social agents required substantial subject matter expertise. Furthermore, most of the categories that performed poorly in the application domain had also shown low performed during k-fold model evaluation (see Table 88). Three classes had an overall low accuracy:

- knowledge, specific, art (rank during k-fold cross validation: 35 (lowest =44))
- organization, specific, product (rank during k-fold cross validation: 43)
- resource, specific, product (rank during k-fold cross validation: 13)

Since these classes were not necessarily needed for further analysis, I removed them from the thesaurus. Also, I removed commas from the retrieved concepts to ensure that the thesaurus complies with the csv format. The quantitative impact of this and all other thesaurus cleaning processes described in this section is summarized in Table 91. However, some of the categories that scored low during cross-validation did not deliver poor results in the application scenario. For example, entries from the category "location, generic, other", which had the lowest performance with class model 4 during cross-validation, returned reasonable results on the Sudan corpus.

Third, many of the erroneous entries originated from the beginning of sentences in the Sudan corpus. Those cases were typically common nouns that would not appear with upper case spelling otherwise. For learning the prediction models, I had included a feature that addressed this situation, and this feature added a meaningful amount of accuracy to the models. Besides potential weaknesses of this feature, there could be other reasons for the observed limitation: the beginning of sentences is also a challenge for the part of speech tagger, which might further lower the certainty with which common nouns are categorized. These problematic cases might further dilute the accuracy of classes where most instances do occur as capitalized tokens at the beginning of sentences and elsewhere, such as specific agents.

Fourth, further screening the thesaurus suggested that some entries differed only in non-letter, non-number symbols, e.g. “NGO” versus “(NGO)”. Other entries resembled delete list entries, i.e. noise terms. To solve these two issues, I identified a list of irrelevant symbols and removed them from all entries while maintaining the remaining content of the impacted entries. Next, I applied the same delete list as used for the Sudan master thesaurus to the auto-generated thesaurus. Only those thesaurus entries that exactly matched a delete list entry were removed.

Fifth, many entities showed up in multiple categories. For example, “muslims” were categorized as agent, generic, noun phrase (frequency = 4) as well as “organization, specific, religious, noun phrase (frequency = 1,276). Like in the given example, many of these alternative assignments are plausible in specific contexts. It depends on the research question and size of the dataset whether one wants to extract these alternative nodes from the texts or not. However, since the thesauri in AutoMap are not yet capable of differentiating between entities of the same class occurring different contexts, I had to remove any alternative categorization for entities with the same surface form (regardless of part of speech). I did that by keeping the entry with the highest observed frequency count. I built and applied a tool that consolidates nodes according to the rules shown in Table 89. Whenever thesaurus entries are merged onto the same concept based on these rules, the frequencies of these entities are added up such that the total cumulative entity frequency remains constant. Table 89: Entity consolidation in auto-generated Funding thesaurus based on matches in certain features

Consolidation based on	**Consolidated if entities match in:**					
	Spelling (case-sensitive)	Meta-network category	Specificity	Subtype	Ratio of unique entities reduced	Ratio of unique entities reduced
POS	x	x	x	x	1.4%	0%
Subtype	x	x	x		3.1%	0%
Specificity	x	x			0.9%	0%
Meta-nw. category	x				10.7%	0%
Word identity					4.6%	5.8%

Reviewing the auto-generated thesaurus at this point suggested that the highly frequent entries seemed correct and no categories with an overall poor performance were still present. However (sixth), inspecting the generated network data in ORA suggested that many entities still occurred in the wrong meta-network category, and with surprisingly high frequencies. Note that this issue was not apparent by reviewing the thesaurus since these entities had a low frequency in the thesaurus, but the respective nodes had high frequency in the network data. For example, “Dr” occurred as “location, specific, country”, but according to the auto-generated thesaurus, should have been categorized as an attribute. Further investigating this issue revealed that AutoMap internally converts every entity in a thesaurus to lower case spelling before translating text terms that match thesaurus entries. This is troublesome for capitonyms, i.e. terms whose meaning differs depending on capitalization: “DR” is a common abbreviation for the Democratic Republic of Congo, and this meaning differs from the meaning of the thesaurus entry “Dr”, which truly is a personal attribute. I found out that if a term appears as capitalized as well as in lower casein the thesaurus, AutoMap by default and without an option to change this behavior picks the thesaurus entry for the term starting with a lower case. Consequently, both “Rice” (the person) and “rice” (the food) are categorized as a resource of subtype substance, and the same is true for “Bush” versus “bush” and “Apple” versus “apple”. Since this AutoMap feature was not up for change, I extended the thesaurus entry consolidation tool described above such that it also merges terms that have the same spelling regardless of capitalization. In this tool, the category assignment of the term with the higher frequency is chosen, and the term frequency is increased accordingly. This consolidation approach improves the status quo of AutoMap since it picks the term – category association with the higher frequency, which might be more appropriate than picking the lower case term by default.

Seventh, further reviewing the thesaurus suggested that the relevance and accuracy of entries drops as their cumulative frequency decreases. More specifically, at low frequencies, entries tend to become long, concatenated chains of multiple relevant entries, e.g. “the Sudan Liberation Movement (SLM) faction of Arkoi Minawi”. Typically, we are interested in representing these entities (in this case “SLM” and “Arkoi Minawi”) as separate ones. Splitting up those chains is also important as AutoMap maps text entries to the longest (in terms of number of tokens) entries it finds in a thesaurus such that long chains will take away matches from shorter entities. Therefore, I removed all entries with a frequency of less than three as three seemed an appropriate cut-off point for this thesaurus.

To further assess the quality of the thesaurus, I reviewed the entity class, specificity value and subtype of all entries with a cumulative frequency of 500 and more (N = 807). These entities account for only 2.09% of all unique entities in the current version of the thesaurus, but for 78.1% of the total entity frequency. I made corrections to the meta-network category, specificity value, or subtype of 39 (4.8%) of these entities. Most of the changes were made to the subtype value, e.g. changing the entities “Doha” and “Eritrea” from “location, specific, city” to “location, specific, country”. This observation (eight) indicates that the small amount of entities that account for the majority of the total entity weight are predicted with high accuracy. Table 90 shows the frequency distribution of these entities.

Table 90: Frequency distribution of entities with cumulative frequency of 1,000 and more in thesaurus*

Class	Thesaurus entries unique	Thesaurus entries total	Average no. of repetitions per entity	Ratio in full thes., unique	Ratio in full thesaurus, total
location, specific	143	786,815	5,502	0.37%	22.19%
agent, generic	233	768,531	3,298	0.60%	21.67%

organization, generic	79	350,351	4,435	0.20%	9.88%
location, generic	38	191,804	5,047	0.10%	5.41%
time	87	171,863	1,975	0.23%	4.85%
attribute	64	153,783	2,403	0.17%	4.34%
attribute, specific	29	122,872	4,237	0.08%	3.47%
organization, specific	65	119,098	1,832	0.17%	3.36%
agent, specific	39	35,927	921	0.10%	1.01%
resource, generic	11	22,146	2,013	0.03%	0.62%
resource	11	21,861	1,987	0.03%	0.62%
knowledge, specific	7	14,260	2,037	0.02%	0.40%
event, specific	1	1,861	1,861	0.00%	0.05%
Total	807	2,761,172	3,422	2.09%	77.87%

* four highest values underlined

Next, I manually reviewed the entries in the categories that I had assessed as having medium or bad performance in the application domain, but were not removed from the thesaurus. I corrected the entries with high frequencies.

At this point, I used the auto-generated thesaurus as part of the D2M process to extracted network data from the texts. I unionized the networks per texts into one network per year, and then the yearly networks into one overall network. In this overall network, I reviewed the highly frequent nodes per meta-network category[16], deleted overly common entities, and made changes to the node-class, specificity value, and subtype were necessary. During this qualitative review, I detected three main types of errors (observation number nine):

- Common nouns that would typically occur in lower case appear as upper case terms; mainly because they are the first word in a sentence. Examples are “Equality” and “Referendum”. This point is consistent with observation number three.
- All letters in common nouns as well as proper nouns are capitalized, e.g. because the term is an abbreviation or the name of an organizations. Examples are WHO (World Health Organization) and LOT (the airline), and TOTAL (the gas company),
- Common nouns as well as word with other part of speech that are typically in lower case are capitalized; mainly because they refer to a named entity with a different meaning. Examples are “Target” (the store) and Nature (the journal).

Instances of all three cases typically occur with a low frequency, and a lower frequency than the more common, lower case version of the those terms. However, since the CASOS tools convert

[16] Entities including and above the following cumulative node frequency values were reviewed: agent, knowledge, location, organization, time, resource: 1,000, event: 100, task: 0. Differences are due to differences in node weight distribution and size of node class; with the “task” class being the smallest.

all entities to lower case when applying thesauri and also compare nodes on a lower-case basis, these outlined special cases cannot be disambiguated via capitalization. Instances of these cases were often predicted as specific agents and organizations, but I corrected many of them by moving them to the knowledge and task classes. Also, I decided to delete all instances of the “organization, specific, other” class with an entity frequency of less than ten since these entries contained too many common nouns. In the future, this problem can be solved by enabling case-sensitivity of the thesaurus application routines and by disambiguation terms based on their part of speech. In fact, both types of information are readily available in the auto-generated thesauri.

Next, I de-duplicated entities again based on surface form and meta-network category. Also, I performed co-reference resolution on the thesaurus by using the same merge lists for nodes from the agent and organization class as developed and used for the network data generated with the Sudan master thesaurus. Table 92 summarizes the frequency distribution of all remaining entities classes across the thesaurus.

Table 91: Summary of thesaurus cleaning routines and quantitative impact

Routine	Entities		Ratio of raw size	
	Unique	Total	Unique	Total
1. Raw auto-generated thesaurus	502,485	5,380,091	100%	100%
2. Remove categories with low performance	283,252	4,115,328	56.4%	76.5%
3. Apply delete list and remove symbols	281,611	3,763,557	56.0%	70.0%
4. Consolidate entries (in named order) based on part of speech, subtype, specificity, meta-network class, spelling regardless of capitalization	227,309	3,763,557	45.2%	70.0%
5. Remove entries with frequency of less than three	38,632	3,546,065	7.7%	65.9%
6. Correct entries with frequency of 500 and more, correct and clean poorly performing categories	38,617	3,537,234	7.7%	65.7%
7. Correct entries after reviewing high frequency nodes in network data, re-deduplicate nodes	35,629	3,480,330	7.1%	64.7%

Table 92: Frequency distribution of entities classes in thesaurus

Class	Ratio in full thes., unique	Ratio in full thesaurus, total	Average number of repetitions per unique entity
agent, specific	24.8%	4.2%	17
attribute	17.0%	7.7%	44
time	15.8%	8.6%	53
location, specific	13.8%	25.1%	178
organization, specific	10.5%	5.7%	53
agent, generic	6.1%	24.2%	388
resource	4.8%	1.5%	30
knowledge, specific	2.3%	0.7%	29

organization, generic	2.1%	11.7%	529
attribute, specific	1.0%	3.8%	374
event, specific	0.6%	0.2%	27
location, generic	0.4%	5.8%	1,324
task, generic	0.3%	0.1%	22
resource, generic	0.2%	0.8%	382
knowledge, generic	0.2%	0.1%	54
resource, specific	0.0%	0.0%	7
Total	100.0%	100.0%	98

* Ratios of 10% and more in full thesaurus underlined

Reviewing the re-generated network data at this point suggested that the thesaurus is sufficiently correct as a stand-along file and for coding texts as networks. I made further refinements to the generated network data files and the attribute files for the networks in ORA directly, such as changing the node class and specificity value of a few nodes, but did not remove any further nodes.

Overall, this section has shown that the network quality improves if the auto-generated thesaurus if verified and corrected, even though this process involves a substantial amount of labor. However, generating and correcting the auto-generated thesaurus is more efficient than building or cleaning a master thesaurus as described in the previous section, where this process took six weeks (5.2.2.1.1). Applying the prediction models for entity extraction takes about one hour per one thousand newspaper articles. Further refining the thesaurus, including building additional post-processing tools and testing various (sequences) of refinement strategies, took about two work weeks. Repeating this process in the future will be more time-efficient as actually shown in the next application case, because parts of this process have now been automated, and a reasonable sequence of steps has been identified and tested.

5.2.2.3 Network Data Construction from Meta Data

Meta-data are a type of structured data that are often available when retrieving news articles from archives such as LexisNexis. In LexisNexis, meta-data are conveniently sorted into categories, e.g. “geographic” and “organization”. Each category can have zero, one or many entities per articles, e.g. “Sudan” and “Khartoum” for “geographic”. Each meta-data entry is associated with a relevance score between zero and one. This score is assigned by LexisNexis without further documentation on this process.

I operationalized link formation between meta-data entities as follows: two entities are linked if they co-occur for the meta-data for an article. This operationalization resembles the notion of windowing such that the network data constructed with the previous two thesaurus-based text coding methods and those built from meta-data are based on the same notion of link formation. Table 93 shows the mapping that I defined for converting LexisNexis meta-data categories into meta-network categories that ORA can interpret.

The output from this process are bidirectional, weighted graphs. The link weights were computed by using a method developed by Pfeffer and Carley (under review), which basically calculates the average of the minima of the relevance scores for the two entities in each link. When the networks per article are merged into consolidated networks – one per calendar year in this case - the cumulative sum of the weight per link is divided by the number of articles in the corpus per year. Thus, all links have a weight between zero and one, but for frequently observed links, this weight has a stronger empirical support, even though this fact is not visible in the network data anymore. The node weight in the aggregated network represents the number of articles that a meta-data entity had been assigned to.

Table 93: Meta-data categories considered, and mapping to meta-network categories

Category in input data	Assigned to meta-network category
Organization	Organization
Company	Organization
Subject	Knowledge
Person	Agent
Geographic	Location

The main advantage with network construction from meta-data is the speed of the process: once the meta-data are downloaded and organized in some structured form, such as a table or a database, generating networks is basically a data retrieval task, which takes a couple of minutes. The limitation with this approach is that the assignment of meta-data entries to articles is not transparent as there is no documentation on what algorithm is used by LexisNexis to generate these index terms and their values.

5.2.2.4 Network Data Construction in Collaboration with Subject Matter Experts

I collaborated with Dr. Richard Lobban, who is s a professor of anthropology and African studies at Rhode Island College (RIC) and a leading expert on Sudan, and his team, notably Adam Gerard and Erica Fontaine, on generating this dataset of tribal affiliations in Sudan. The RIC team had provided us with a list of the main tribes in the Sudan. I applied this list as attributes to network data that I had previously generated by using the standard data coding process in AutoMap as described in 5.2.2.1 (with master thesauri) such that some organizations were cross-classified as tribes. Then, I used ORA to extract the sub-network of tribes and generated a network visualization of the tribal affiliation network per calendar year. I sent these network visualizations to Dr. Lobban's team, who then marked up the missing nodes and links (false negatives) and invalid nodes and links (false positives). They scanned their maps and sent them back to me, and I made the respective changes to the DyNetML files. We repeated this process until Dr. Lobban's teams considered the networks as representative of the ground truth according to their subject matter expertise.

The advantage with this process is that it results in validated network data, which is the only relational ground truth data that I have available for the Sudan. However, there are also two disadvantages with this network data construction method: first, this process is expensive in terms of time and human resources; going through this process took several weeks. This amount of time is comparable to what is needed for constructing or cleaning thesauri. However, in contrast to both thesaurus construction methods in AutoMap, further time reductions for this process are unlikely since these domain specific data cannot be expected to generalize to other domains, while the thesauri might contain entries of general interest that might also occur in other domains. Second, this process does not scale up, and is therefore only appropriate for generating datasets of small to moderate size.

5.2.3 Results

The frequency distributions of the identified entities per class suggest two findings (Table 90, Table 92): first, all of the classes that I evaluated as having "medium" or "bad" prediction quality in the application scenario have the value "specific" for the specificity attribute. Second, the vast majority of all retrieved entities as well as of entities with a frequency of 1,000 or more, which I manually evaluated as being classified correctly to 96.8%, have the specificity value "generic" – with the exception of the class "location", where specific instances are more prevalent than generic ones. Taking these observations together, I argue that even though network analysis is often focused on analyzing nodes that represent named social entities; i.e. individual people and groups, most of the potential nodes contained in text data are references to social collectives, e.g. types and roles of people and groups. From this finding, I conclude that understanding the impact of collectives on networks, their participants and wider contexts requires not only a) performing analysis on the level of social roles, but also b) considering unnamed entities in addition to named entities in the first place. However, data on these unnamed entities is often not collected with traditional network data collection methods. Therefore, I argue that using entity extraction from text with the approach and technology developed, implemented and evaluated in a lab study and applications scenarios in this thesis can offer a highly valuable addition to classic network data collection methods.

In the following, I refer to network data generated with master thesauri as “D2M”, to network data constructed with the auto-generated thesauri as “D2M+EE”, to networks data built from meta-data as “META”, and to network data generated in collaboration with subject matter experts as “SME”. Reported averages were computed across the networks per year; excluding the union graph, unless specified otherwise.

The size of the networks depending on the network data construction method (Table 94, Table 95) show that even though the auto-generated thesaurus (D2M+EE) is 4.8 times smaller than the master-thesaurus (D2M), the D2M+EE networks have on average about 1.5 more nodes and 1.7 more edges than the D2M networks. Furthermore, 11.5% of the entities contained in the master thesaurus (N=19,489) (D2M) occur in the D2M+EE networks, while 72.4% of the entities contained in the auto-generated thesaurus (N=25,794) (D2M+EE) appear in the D2M networks (Table 94). These results together with the implied conclusion that the auto-generated thesaurus (D2M+EE) leads to more matches in the text data while having less entries than the master thesaurus (D2M) suggests that the auto-generated thesaurus is more effective than the master thesaurus in covering the dataset and domain. However, from a practical point of view, the rate of entities that are specified in the thesaurus but are not contained in the data is mainly irrelevant: non-matching entries are disregarded, which has a minor impact on runtime. In summary, since the master thesaurus took three times longer (six weeks) to generate and post-process than the auto-generated thesaurus (two weeks), using the auto-generated thesaurus for text coding as part of the D2M process (D2M+EE) seems more efficient and effective than working with master thesauri (D2M).

Both types of networks extracted from the text bodies (D2M, D2M+EE) are larger than the meta-data networks in terms of nodes (D2M: 2.5 times larger, D2M+EE: 3.8), and for the D2M+EE networks also in terms of links (D2M+EE: 1.4, D2M: 0.8).

In chapter 2.7.2 of this thesis, I had shown that the windowing approach to link identification, which has been used in this application scenario, can lead to a significant amount of false positive links. The networks built from text bodies are subject to this source of error. However, if we assume that the meta-data networks serve as a point for reference for the number of links or graph density, the difference in the amount of links between the meta-data networks and text-based networks is more than three times smaller than the difference in the amount of nodes. The counterargument to this point is that the meta-data networks were also constructed based on co-occurrence; an approach which is resembled in the windowing approach.

In the previous methods section I had shown that not only the master thesaurus, but also the auto-generated thesaurus need further manual and computer-supported cleaning to correct for misclassified entries and remove overly generic suggestions. Table 94 shows that the number of nodes and edges that get removed due to this process is very similar across the annual networks (1.6% difference). This observation indicates that the number of links does not shrink slower than the number of nodes, which further relates to the potential amount of false positive links, and also suggests a reduced likelihood of this risk. However, it is unclear if the same trend holds for the opposite direction, i.e. if the number of links grows faster than the number of added nodes depending on the network construction method or not. This relationship is beyond the scope of this thesis, but should be addressed in future work.

Table 94: Network size per network construction method I

Data	SME		D2M		D2M with EE		Meta-data		Articles
	Nodes	Links	Nodes	Links	Nodes	Links	Nodes	Links	per year
Thes. entries	n.a.		169,791		35,629		n.a.		n.a.
2003	21	15	6,612	142,630	9,932	221,104	4,648	203,274	4,507
2004	26	22	9,894	288,051	14,750	483,862	7,093	441,076	10,059
2005	22	15	9,420	258,502	14,189	434,525	5,765	381,732	7,837
2006	23	27	10,837	345,796	16,313	600,748	3,677	421,896	11,076
2007	23	40	11,195	360,886	16,876	619,204	3,897	465,378	12,243
2008	36	50	10,303	318,721	15,920	539,559	3,374	377,652	10,713
2009	n.a.	n.a.	9,537	294,344	15,024	496,961	2,986	312,228	10,410
2010	n.a.	n.a.	9,378	304,659	15,315	527,851	2,931	294,928	12,543
Union Graph	53	104	19,489	1,130,934	25,794	2,296,397	15,128	1,561,528	79,388

Table 95: Network size per network construction method II

Category	SME	D2M	D2M + EE	Meta-data
Number of node classes	1	8	8	4
Number of networks	1	36	36	16

Table 96: Network size depending on thesaurus cleaning

Data	Raw		Post-processed thes. (step 7)		Ratio of reduced to raw	
	Nodes	Edges	Nodes	Edges	Nodes	Edges
Thes. entries	502,485		35,629			
2003	20,393	498,593	9,932	221,104	48.7%	44.3%
2004	35,092	1,228,551	14,750	483,862	42.0%	39.4%
2005	33,950	1,073,384	14,189	434,525	41.8%	40.5%
2006	41,569	1,448,364	16,313	600,748	39.2%	41.5%
2007	43,994	1,550,240	16,876	619,204	38.4%	39.9%
2008	39,384	1,317,270	15,920	539,559	40.4%	41.0%
2009	36,576	1,204,194	15,024	496,961	41.1%	41.3%
2010	39,791	1,378,412	15,315	527,851	38.5%	38.3%

Union graph	134,507	6,194,467	25,794	2,296,397	19.2%	37.1%

How similar are the networks per network data construction method to each other on a structural level? I answer this research question by a) computing the intersection of any pair of networks per year as well as of the unionized graphs and b) calculating the amount of nodes and edges from any one network that are also present in any network constructed with another method for the same time period. The results from intersecting the SME networks, which can be considered as ground truth data, with the D2M networks show that over half of the nodes and a fifth of the links are present in D2M (Table 97). Also, the D2M networks resemble 2.6 times more of the nodes and 3.7 times more of the edges from the SME network than the D2M+EE networks (Table 97). I assume this outcome to be due to the fact that a list of tribes in the Sudan (the nodes in the SME network) that our project partners at ROC and ECU identified was also added to the master thesaurus (D2M), but not to the auto-generated thesaurus (D2M+EE). In contrast to that, all of the tribes listed in the auto-generated thesaurus as specific organizations were identified by the entity prediction models based on the content of the text data only. Moreover, the intersection between the SME networks and the meta-data networks is zero on the node and link level (Table 97).

Table 97: Resemblance of ground truth data per network construction method

Data	SME contained in D2M		SME contained in D2M+EE	
	Nodes	Links	Nodes	Links
Thes. entries				
2003	52.4%	13.3%	23.8%	6.7%
2004	46.2%	40.9%	23.1%	9.1%
2005	63.6%	33.3%	27.3%	20.0%
2006	47.8%	33.3%	21.7%	7.4%
2007	78.3%	12.5%	26.1%	5.0%
2008	41.7%	28.0%	11.1%	4.0%
2009	n.a.	n.a.	n.a.	n.a.
2010	n.a.	n.a.	n.a.	n.a.
Union Graph	52.8%	20.2%	11.3%	4.8%

Disregarding the SME network, the intersections between the remaining types of networks are strongest between D2M and D2M+EE; with D2M+EE resembling twice as much of D2M than vice versa (Table 98). Overlaps between the networks derived from texts with meta-networks are small: the text-based networks pick up only a small amount of the nodes contained in the meta-networks (7.8% - 11.5%), and hardly any of their links (less than 1.2%). The meta-networks contain less than 5.2% of the nodes in the networks derived from texts, and less than 1.2% of their links. Overall, the network size seems to impact the mutual resemblance of networks: the larger a network, the higher the chance that constituents from another network are also contained therein.

Table 98: Intersection of nodes and links per year and method

Data	Intersection of D2M and D2M+EE				Intersection of D2M and Meta-data				Intersection of D2M+EE and Meta-data			
	D2M+EE contained in D2M		D2M contained in D2M+EE		Meta-data contained in D2M		D2M contained in Meta-data		Meta-data contained in D2M+EE		D2M+EE contained in Meta-data	
	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges
2003	15.0%	5.0%	22.5%	7.7%	8.5%	0.2%	5.9%	0.2%	6.8%	1.2%	3.2%	1.1%
2004	13.5%	4.7%	20.1%	7.9%	11.3%	0.2%	8.1%	0.2%	7.1%	0.9%	3.4%	0.8%
2005	13.8%	4.8%	20.8%	8.1%	12.0%	0.2%	7.4%	0.3%	8.1%	1.1%	3.3%	1.0%
2006	13.1%	4.8%	19.6%	8.4%	13.4%	0.2%	4.5%	0.2%	8.1%	1.2%	1.8%	0.8%
2007	12.7%	4.8%	19.2%	8.3%	12.7%	0.2%	4.4%	0.2%	7.5%	1.1%	1.7%	0.9%
2008	12.7%	4.9%	19.7%	8.3%	12.2%	0.2%	4.0%	0.2%	8.0%	1.2%	1.7%	0.9%
2009	12.9%	4.8%	20.3%	8.1%	12.1%	0.2%	3.8%	0.2%	8.4%	1.2%	1.7%	0.8%
2010	12.4%	4.8%	20.2%	8.3%	10.2%	0.2%	3.2%	0.2%	8.2%	1.2%	1.6%	0.7%
Union	10.4%	4.4%	13.7%	8.9%	11.6%	0.2%	9.0%	0.3%	5.5%	0.9%	3.2%	0.6%
Average (years)	13.3%	4.8%	20.3%	8.2%	11.5%	0.17%	5.2%	0.22%	7.8%	1.1%	2.3%	0.9%
Rank nodes	2		1		3		5		4		6	
Rank links		2		1		5		6		3		4

Another important question for practical applications is whether it is worth the effort to clean auto-generated thesauri or not. The results show that using the auto-generated thesaurus as is to generate D2M+EE networks results in the retrieval of less than half the amount of nodes (48.4% for D2M, 48.5% for meta-data) and only a small fraction of the links (3.0% for D2M, 0.1% for meta-data) in comparison to network data generated with the refined auto-generated thesaurus (Table 99). This means that with only 14.1% of the thesaurus entries left; many of which had been subject to correction (Table 91), more than twice as many nodes are found in the intersection, and also the vast majority of links is only retrieved after this cleaning process. This finding further emphasizes my previous conclusion that post-processing the output from the entity prediction models is crucial and unavoidable.

Table 99: Impact of refinement of auto-generated thesaurus on network intersection

Data	Ratio of final D2M+EE intersection with D2M contained in intersection of D2M+EE (raw auto-generated thesaurus) and D2M	Ratio of final D2M+EE intersection with meta-data contained in intersection of D2M+EE (raw auto-generated thesaurus) and Meta-data

	Nodes	Edges	Nodes	Edges
2003	39.4%	2.7%	63.5%	0.1%
2004	49.8%	3.1%	70.8%	0.2%
2005	46.0%	2.7%	62.7%	0.3%
2006	50.6%	3.0%	39.7%	0.1%
2007	53.7%	3.4%	48.5%	0.1%
2008	50.4%	3.0%	40.5%	0.0%
2009	48.3%	3.2%	30.8%	0.1%
2010	49.0%	3.4%	31.5%	0.1%
Union	88.1%	4.1%	115.4%	0.3%

To further compare the networks per construction method, on a very general level, one can choose between computing network metrics on the data or identifying key entities in the data, among other methods. For this chapter, I made this choice based on the insights gained in the previous chapters: the master thesauri used in this chapter, and to a lesser degree also the auto-generated thesauri, have been subject to semi-automated as well as manual co-reference resolution. I conducted this co-reference resolution for each thesaurus separately, but reused material such as node merger list within and across the application scenarios. Based on the experimental results from chapter 2 and the practical implications of these results described in chapter 4, conducting reference resolution is essential for extracting entities from text data. However, since AutoMap does not yet offer a sufficiently accurate anaphora resolution routine and no automated co-reference resolution routine, I only performed co-reference resolution on the thesauri in a mainly and computer-supported fashion with the tools developed and described in this chapter. Consequently, the values of network metrics computed on these networks can be expected to be less accurate in terms of resembling the ground truth than key entities identified from these data. This is because I have shown that key entities are less sensitive to variations in network size and imperfect reference resolution techniques than network metrics (chapter 2). Thus, key entity analysis is a more reliable strategy for analyzing and contrasting the network data than network metrics would be. Therefore, the key entity analysis method is used throughout this chapter.

The results for network overlaps on the structural level as presented earlier in this chapter had suggested that the meta-networks represent a different set of information than the text-based networks. Does this also hold true for the prominent nodes in the network? In other words, how similar are the networks per network construction method to each other on a qualitative level? I answer this research question by conducting key entity analysis as follows: I partitioned the networks so that for the classes of agents and organizations, only specific instances are kept. Next, I identified the top 15 entities per network construction method (D2M, D2M+EE, meta-data), network analysis metrics (degree centrality, betweenness centrality, eigenvector centrality, clique count, for a definition of the metrics see Table 154), node type (agent, organization, knowledge), and calendar year (2003-2010). I output this data over-time, ranked the top entities per network type, node class, and metric, and computed the average rank per entity over the considered time period. If an entity did not show up in one or more of the years considered, I assigned rank number 15 (the lowest possible rank) to it. I chose this method for identifying key players from over-time data because it jointly considers continuity *and* prominence of an entity, and also makes this kind of three-dimensional information (overtime, across methods, across entities) representable in Table form. Finally, I performed manual co-reference resolution on the key players per network type: I screened the top 15 entities for text-based and meta-data networks, and converted different spellings of entities that most likely refer to the same real-world entity to the same surface form, e.g. “bush” and “george bush” to “bush”, or “talks and meetings” and “meetings” to talks & meetings”.

The results from the key player analysis show that between D2M and D2M+EE, there is a substantial overlap in agents, and to a lesser degree also in organizations (Table 100 to Table 104). For example, across the four network metrics considered on the agent level, D2M and D2M+EE share 55% of the key agents. The agreements are lower on the organizations level. In the text based networks, most of the key agents are Sudanese politicians, but a few international and other African individuals are also highly prominent, e.g. “Yoweri Museveni”, the president of neighboring Uganda. Most of the key organizations are political and governmental units as well as instances of armed forces, including rebel groups such as the Janjaweed and the Lord’s Resistance Army. Again, most of them are Sudanese, but the key organizations include more international entities than the key agents, mainly groups from the USA and the United Nations, such as the “International Criminal Court”, which had issued warrants for multiple Sudanese politicians, mainly because of their involvement in the Darfur conflict.

The comparison of entity results between D2M and D2M+EE also suggest that considering the content of the text bodies leads to the retrieval of highly central first names, such as “Muhammad” and “Ahmad” (these names are marked in gray font in Table 100). While it might be reasonable to consolidate “Joseph” and “Kony” (Joseph Kony is the leader of the Lord's Resistance Army), such names cannot necessarily be mapped onto single individuals: for example, “Muhammad” or “Ahmad”, which could refer to the two distinct individuals “Ahmad Al-Bashir” or “Muhammad Ahmad”, who are both prominent figures in the given domain, such a mapping would be more speculative, might result in picking up on false positives, and would require substantial subject matter expertise to make this judgment. The meta-data networks do not have this issue, but are also not free of entity disambiguation issues. For instance, in META, Omar al-Bashir occurs as “Omar Hassan Ahmad al-Bashir” as well as “Omar al-Bashir”.

The overlaps between the meta-data networks and the text-based networks are smaller than the overlaps between the text-based networks. Also, between the meta-data networks and text-based networks, the overlap is larger with respect to key organizations than key agents. In fact, the text-based networks and meta-data networks only agree on two key agents, namely Al-Bashir and George Bush. For organizations, the intersection is about equally split up among Sudanese and foreign or international organizations. However, most of the key organizations in the meta-networks are non-Sudanese groups, but in contrast to the text-based networks, they include groups from industry and also a large portion of international NGOs. The key individuals in the meta-data networks are mainly high-profile, international politicians, such as Hillary Clinton and Ban Ki-Moon, and other prominent international figures involved in politics, such as George Clooney, who has actively promoted the development of the Sudan. Further looking into the data revealed that many of these entities occur in the same node classes in the text-based networks, but with lower prominence.

Table 100: Key agents per network construction method and metric I*

Degree Centrality				**Betweenness Centrality**			
Key entity	D2M	D2M+EE	Meta-data	Key entity	D2M	D2M+EE	Meta-data
al-bashir	1.6	1.6	5.3	garang	1.5	1.4	
taha	1.9	2.3		al-bashir	1.9	2.9	5.3
muhammad	3.9	3.9		taha	3.4		
ahmad	5.4	9.9		bush	6.5	6.0	4.3
garang	6.4	6.1		muhammad	6.5	7.1	
ibrahim	7.6	10.6		ahmad	7.6	11.3	
hassan	8.3	9.9		ibrahim	9.0	10.5	
bush	9.1	10.9	1.8	deng	9.0	11.6	
kony	9.3	7.1		ahmed	9.9		
kiir	9.8	9.0		david	9.9		
ahmed	10.3			adam	10.0		
joseph	10.3			joseph	10.8		
ismail	11.5			michael	11.0	10.0	
abdallah	12.3			kiir	11.3		
mohamed	12.4			ismail	11.9		
ali		3.8		kony		5.0	
museveni		10.4		ali		6.0	
mustafa		10.5		james		7.8	
annan		12.0		paul		8.1	
isma		12.0		george		10.1	
				museveni		10.5	
				peter		11.8	
hillary_rodham_clinton			6.3	tony_blair			6.8
tony_blair			7.0	hillary_rodham_clinton			7.3
bill_clinton			7.3	barack_obama			7.5

michael_mcmahon	7.5	michael_mcmahon	7.6
condoleezza_rice	8.1	condoleezza_rice	8.0
ban_ki-moon	8.6	ban_ki-moon	8.4
barack_obama	8.6	osama_bin_laden	9.1
thabo_mbeki	8.9	george_clooney	9.5
tzipora_livni	8.9	mahmoud_ahmadinejad	9.9
gordon_brown	10.4	saddam_hussein	10.1
hu_jintao	10.9	gordon_brown	10.3
nicolas_sarkozy	10.9	nicolas_sarkozy	11.1
george_clooney	11.6	hosni_mubarak	12.5

* First names that may refer to multiple people grayed out in this table.

Table 101: Key agents per network construction method and metric II

Eigenvector Centrality				**Clique Count**			
Key entity	D2M	D2M+EE	Meta-data	Key entity	D2M	D2M+EE	Meta-data
al-bashir	2.5	2.0	6.3	al-bashir	1.4	1.1	5.9
taha	3.1	5.0		taha	1.6		
hassan	5.5	4.6		muhammad	4.0	3.3	
muhammad	5.6	8.0		ahmad	6.4	6.6	
ahmad	7.0	8.6		ibrahim	6.8	6.9	
kiir	7.1	7.4		garang	6.9	3.9	
garang	7.8	8.6		ahmed	6.9		
museveni	8.5			adam	8.5		
ismail	9.1			abdallah	10.0		
ibrahim	9.6			bush	10.1	8.0	1.6
kony	10.3			mohamed	10.1		
mustafa	10.4	10.6		hassan	10.9		
abdallah	10.5			ismail	10.9		
osman	11.0			mohammed	11.9		
joseph	12.0			musa	13.1		
hasan		6.4		ali		3.5	
ali		6.9		kony		7.4	
republic_field_marshal_umar		8.6		deng		9.8	
deby		9.4		museveni		10.1	
annan		10.0		james		10.9	
isma		11.3		paul		11.1	
powell		12.6		george		11.4	
				peter		13.6	
				michael		13.8	
bush			2.3	condoleezza_rice			6.5
hillary_rodham_clinton			6.9	saddam_hussein			8.5
tony_blair			7.6	nicolas_sarkozy			9.0
condoleezza_rice			8.0	tzipora_livni			9.0
bill_clinton			8.3	mahmoud_abbas			9.1
saddam_hussein			8.8	vladimir_putin			9.3

barack_obama	8.9	michael_mcmahon	9.4
ban_ki-moon	9.3	tony_blair	9.5
tzipora_livni	9.5	ban_ki-moon	11.0
osama_bin_laden	10.1	angela_merkel	11.8
michael_mcmahon	10.3	ehud_olmert	11.8
thabo_mbeki	11.0	mahmoud_ahmadinejad	12.3
robert_zoellick	11.6	barack_obama	12.9
j_scott_gration	11.8		

Table 102: Key organizations per network construction method and metric I

Degree Centrality				**Betweenness Centrality**			
Key entity	D2M	D2M +EE	Meta-data	Key entity	D2M	D2M +EE	Meta-data
government	1.0			government	1.0		
forces	2.5			forces	2.4		
spla_splm	3.5	2.6	12.8	military	3.3	5.0	
military	3.9	8.8		national_council	4.3		
us_army	6.3			spla_splm	4.9	6.4	
national_council	8.0			us_army	8.5		
lords_resistance_army	8.8	5.6	12.0	police	9.0		
janjaweed	9.8	11.9		us_congress	9.6		
united_nations	10.1	1.8	1.3	sudan_embassy	9.8		
african_union	10.3	4.6	3.6	united_nations	10.4	1.5	1.1
police	10.4			ruling_party	10.4		
sudan_embassy	10.8			dinka	11.1		
ncp	11.6	10.8		non_gov._organization	11.4		
internat._criminal_court	11.6	12.1	6.8	european_union	12.0		7.5
jem	11.6			foreign_company	12.1		
security		3.3		security		1.9	
army		6.3		southern_sudan		6.5	
humanitarian		8.5		african_union		6.8	4.1
southern_sudan		9.3		humanitarian		7.4	
party		11.0		party		7.6	
militia		11.4		army		7.8	
defense		12.3		defense		9.0	
				the_sudanese_government		11.4	
				justice		11.8	
				opposition		12.0	
				services		12.5	
				university		12.6	
united_nations_security_council			4.3	internat._criminal_court			6.3
european_union			5.5	united_nations_security_council			7.0

league_of_arab_states	8.5	al-qaeda	7.4
human_rights_watch	8.6	african_development_bank_group	10.1
united_nations_world_food_programme	9.6	united_nations_world_food_programme	10.1
liberation_movement	10.9	cninsure_inc	10.3
Intergov._authority_on_development	11.1	sudanese_tv	10.3
united_nations_children_fund	11.3	east_african_community	10.5
sudanese_tv	13.0	human_rights_watch	10.9
inter-governmental_authority	13.6	united_nations_children_fund	11.5
		china_national_petroleum_corp	11.8
		liberation_movement	13.1

Table 103: Key organizations per network construction method and metric II

Eigenvector Centrality				**Clique Count**			
Key entity	D2M	D2M +EE	Meta -data	Key entity	D2M	D2M +EE	Meta -data
government	1.0			government	1.0		
forces	2.5			military	3.0	3.9	
military	3.9	7.4		forces	3.0		
spla_splm	4.0	4.8		national_council	4.4		
us_army	6.3			spla_splm	4.5		14.1
janjaweed	7.5			sudan_embassy	7.1		
lords_resistance_army	8.5	7.8		united_nations	7.5	1.9	1.1
police	9.6			us_army	8.4		
sudan_embassy	9.6			police	9.9		
national_council	10.3			internat_criminal_court	10.9	12.3	13.4
Justice&equality_movemt	10.9			lords_resistance_army	11.1	11.5	
goss	11.0			ruling_party	11.3		
african_union	11.3	5.3	3.9	un_security_council	12.5		4.5
rebel_groups	11.5			ncp	12.6		
ncp	12.3	11.0		us_congress	12.6		
united_nations		3.5	1.5	security		1.6	
security		4.0		splm		4.4	
army		6.8		army		6.8	
humanitarian		7.0		southern_sudan		6.8	
southern_sudan		9.4		humanitarian		7.5	
the_sudanese_government		9.4		african_union		8.1	3.8
party		10.4		party		10.4	
assembly		10.8		justice		10.6	
sudan_peoples_liberation_movem.		11.0		defense		11.0	
internat._criminal_court		11.8	6.8	the_sudanese_government		11.6	
				opposition		11.8	
european_union			7.8	european_union			4.9

human_rights_watch	8.1	united_nations_world_food_programme	8.6
united_nations_world_food_programme	9.8	united_nations_childrens_fund	9.3
league_of_arab_states	10.6	cninsure_inc	10.9
united_nations_security_council	10.9	liberation_movement	11.3
cninsure_inc	11.0	human_rights_watch	13.3
liberation_movement	12.3	sudanese_tv	13.3
arab_league	12.8	talks_&_meetings	13.6
sudanese_tv	12.9	al-qaeda	13.8
united_nations_childrens_fund	13.4	security_council	15.0
inter-governmental_authority	15.0		
security_council	15.0		

In contrast to the social agents level, the text-based networks show no agreement in knowledge nodes, but a small overlap each (about two nodes) with the knowledge nodes in the meta-data networks (Table 104, Table 105). However, the key knowledge nodes in META contain entities that are classified as generic agents and organizations in the text-based network data, e.g. “refugees” and “displaced persons”. Therefore, the overlap between the meta-data networks and text-based networks might be larger if further adjustments were made to the meta-data. In D2M, the key knowledge nodes seem to represent a variety of topics, some of which are highly general, e.g. “political” and “emotion”. This is because almost all of the key knowledge nodes in D2M originated from the RER-cross classification. In contrast to that, D2M+EE and META center on negotiations between political parties and legislative issues, e.g. the Comprehensive Peace Agreement, and also economic issues (D2M+EE), e.g. “trade”.

Table 104: Key knowledge nodes per network construction method and metric

Degree Centrality				**Betweenness Centrality**			
Key entity	D2M	D2M +EE	Meta-data	Key entity	D2M	D2M +EE	Meta-data
peace_process	1.0		8.6	peace_process	1.3		7.6
conflict_knowledge	2.0			time	3.3		
time	3.0			war_&_conflict	3.4		8.0
economy	5.0			literature	7.4		
security_forces	5.0			political_democratizat.	7.8		
political_democratizat.	6.1			measures_numerology	7.9		
valence_pos	6.9			valence_pos	7.9		
emotion	9.8			economy	9.0		
measures_numerology	10.4			political	9.4		

war_&_conflict	10.5		5.6	ideology	10.1		
political	11.6			war	10.1		
biomass_&_land_cover	11.8			communication	10.3		
health	12.1			sovereignty	10.5		
political_displaced	12.1			acknowledgement	10.8		
sovereignty	12.8			security_forces	11.1		
treaties_&_agreements		1.6	11.0	treaties_&_agreements		1.3	6.4
cpa		4.6		cpa		5.1	
sharing		6.1		bill		6.0	
relations		6.6		relations		6.4	
english		6.9		leading		6.5	
summit		7.5		summit		8.0	
trade		7.6		speech		8.6	
website		7.9		website		9.0	
wealth		8.1		policy		9.1	
framework		8.5		talks_&_meetings		9.4	9.1
constitution		9.8		release		9.6	
solution		10.5		constitution		10.0	
musa		10.9		peace_agreement		10.1	
education		11.0		trade		10.3	
industry		13.1		accord		11.5	
international_relations			1.6	religion			1.6
talks_&_meetings			3.9	international_relations			3.5
united_nations_institutions			5.0	refugees			4.8
rebellions_&_insurgencies			7.8	muslims_&_islam			11.0
state_departments_&_foreign_services			9.4	united_nations_institutions			11.0
displaced_persons			11.0	children			11.5
peacekeeping			11.5	armed_forces			12.1
relief_organizations			11.6	rebellions_&_insurgencies			12.4
international_law			12.5	legislative_bodies			12.5
refugees			13.6	international_assistance			13.1
paramilitary_&_militia			14.5	terrorism			14.9

5.3 Application Context II: Funding Corpus

Some federal funding agencies are obligated to publicize information about the allocation of tax-dollars to people, organizations and ideas. For example, the National Science Foundation (NSF) provides a database with information on all previously funded research projects (NSF). The availability of such data has contributed to the transparency of state-level decision making processes. Furthermore, these data allow for addressing substantive research questions such as:

- Business perspective: What team configurations (institutions, disciplines, nationality, gender, …) have been successful in acquiring funding? How does funding impact team dynamics? (Biocca & Biocca, 2002; Horta, Huisman, & Heitor, 2008)
- Social networks perspective: Which individuals and/ or organizations have been collaborating on what? What is the impact of funding research topics on the advancement of a discipline? (Folkstad & Hayne, 2011; Leung, 2007; Melkers & Wu, 2009)
- Human computer interaction perspective: Under what conditions are teams involved in collaborative work sustaining or changing? (Cummings & Kiesler, 2005)

5.3.1 Data[17]

Information about the research proposals that have been accepted and funded through the "Framework Programmes for Research and Technological Development", short Framework Programmes (FPs), is publically available from the Community Research and Development Information Service (CORDIS). The FPs are funded by European Union (EU). The EU Research Council started the first FP in 1984 with the goal of stimulating and enabling competitive research in the European Research Area. The FPs have been continued since then, with the 7th FP currently under way. I used the following process to collect and normalize the Funding corpus:

For this study, I define a "project" as a CORDIS database entry for which at least a unique identification number is provided. Based on this definition, CORDIS contains 55,972 projects for FPs 1 through 6 as of December 2009. I downloaded these data into a relational database, where I performed further data management and cleaning routines. CORDIS provides the projects' start and end date, costs and amount of funding awarded, completion status, and various key words and index terms; all of which I added into my database.

Per project, CORDIS also specifies the name, affiliation and contact information for the project coordinator (PC). PCs are equivalent to principal investigators in the US. The same information is given for each collaborator on a project if applicable. I define a "project with PC" as a project for which a valid entry for the project coordinator is available. An entry is considered as valid if it does not contain any phrase from a set of phrases[18] that I identified by manually going through the people listed in CORDIS.

[17] Portions of this section and the next chapter are reprinted, with permission, from: Diesner, J., & Carley, K. 2010). A methodology for integrating network theory and topic modeling and its application to innovation diffusion. Proceedings of IEEE International Conference on Social Computing (SocComp), Workshop on Finding Synergies Between Texts and Networks, Minneapolis, MN.

[18] These entries are: N/A N/A (N/A), N/A N/A, N/A, NOT AVAILABLE, NOT AVAILABLEE, Address, TBC, tbc TBC, F3 A3.

The project entries also comprise three fields of unstructured, natural language text data: a title, description ("objective") and additional information per project. The length of the text data per project varies greatly; ranging from concise summaries spanning a few sentences to elaborated descriptions. I define a "project with text" as a project for which the length of the project description plus general information exceeds a minimum length of ninety characters after disregarded certain phrases[19]. The minimum length criterion was established to discount for text fields that contain nothing but a generic header, such as "Research objectives and content:". The set of phrases to disregard are expressions that I assessed as being highly common yet not content-bearing in the context of the dataset. Some of these phrases might be parts of the proposal template.

Similar to performing co-reference resolution on the Sudan thesauri, one major challenge with this dataset was the consolidation of the various instances and spellings of people's names into one consistent name per actual individuals. The findings from chapter 2 have shown that high accuracy in this step is crucial because errors from reference resolution of names get propagated to the link and network data level, where they cause biases in the network structure and analysis results. In order to identify the various references to a person, I developed a data-driven set of rules and heuristics, which I iteratively applied and evaluated for their effectiveness and correctness by manually checking their impact on the data: first, all gender and role identifiers, such as "Mrs." and "Professor" were removed from the names. Single-letter umlauts were converted into the equivalent diphthong. All tuples of identically spelled names were considered to represent the same person if their institutional affiliation and/or their address matched completely or at least in three consecutive tokens. Here, tokens are any combination of space separated letters and/or digits. The word "the" was disregarded from this process. People without a valid name entry were also disregarded. In total, my database contained 293,974 entries in the person field. Of those entries, 74.9% were valid people entries. Of those valid entries, 65.2% were identified as unique people (N = 143,700); the others are additional occurrences of those unique people.

At this point, we inspected the resulting database and decided that the procedures that I had developed and implemented for the purpose of data normalization, cleaning and co-reference resolution seemed sufficient. Overall, the completeness of project entries in CORDIS varies per FPs; with later programmes being more complete. Table 105 provides an overview of the size and completeness of the CORDIS database per FP.

[19] The disregarded phrases are: APPROACH AND METHODS, Brief description, Objectives and content, PROJECT DESCRIPTION, Project Details, PROJECT OBJECTIVES, Research objectives and content, Summary of the project, Technical Approach

In this study, I consider data from FP1 to FP6 only; disregarding the downloaded information for FP7. The reason for this decision is that entries for FP7 are still being added, so that the data for FP7 would be incomplete. This is problematic as it has been previously shown that incomplete network data can lead to strongly biased analysis results (Borgatti et al., 2006; Frantz et al., 2009). However, any hypotheses gained from this study can be tested in the future with data from FP7. The same issue with incomplete network data also applies to FPs 1-3, where the ratio of projects with a person is less than 80%. For FPs 4-6, this ratio exceeds 80%, which is considered an acceptable rate for social network data.

Table 105: Size and completeness of research funding dataset

FP Number	Time frame	Number of projects	Projects with text	Projects with PC	Projects with text and PC	Number of unique people	Total number of people mentions	Average agent node weight
1	1984–1987	3,283	82.7%	77.0%	69.8%	2,404	3,246	1.4
2	1987–1991	3,884	79.9%	61.8%	56.8%	6,538	8,544	1.3
3	1991–1994	5,529	76.8%	64.8%	60.1%	14,970	18,407	1.2
4	1994–1998	15,061	79.9%	82.2%	64.1%	37,344	58,682	1.6
5	1998–2002	17,629	75.3%	95.0%	71.9%	36,420	75,355	2.1
6	2002–2006	10,586	96.8%	89.5%	86.8%	43,530	56,066	1.3

5.3.2 Network Data Construction Methods

I used the same methods for generating network data from the Funding corpus as I used for the Sudan corpus where possible to enhance the comparability and generalizability of my findings. In this chapter, I only use data on projects for which at least one PI and a text entry are available (projects with text and person), because both elements are relevant for testing the network agreement in structure and key entities. One limitation with the Funding corpus is that we do not have any ground truth data from subject matter experts on who collaborated with whom. However, one could argue that the social network data about collaborators is highly accurate even though it is self-reported by the respective PIs. Therefore, the social network data created from the meta-data can be considered as ground truth data. The same argument could be made for knowledge meta-networks built from the predefined as well as self-defined index terms that the PIs have selected for their projects.

5.3.2.1 Network Data Extraction from Texts Using the Data to Model Process

The key component for the D2M process is a thesaurus. However, since the master thesaurus built for the Sudan project cannot be expected to generalize well to the domain of research and science, I built a new, domain specific thesaurus (Funding master thesaurus) for the Funding data as follows: first, I worked through the standard D2M process for creating a thesaurus and integrating it with other thesauri: I applied the same delete list as used for the Sudan project to the Funding corpus. Second, I used AutoMap to compute the absolute and weighted (as per tf*idf) frequency per token, and also a list of bigrams per project. AutoMap outputs this information, but it is up to the user to select the appropriate entries. I reviewed the top 550 entries from the frequency lists and the top 1,000 entries from the bigram list (relevance of entries seemed to drop from those frequencies on), and added the concepts that I considered as relevant to the thesaurus (about 1,000). Third, I enhanced the thesaurus with meta-data from CORDIS, which is an example of a domain thesaurus (about 3,000 entries): I used the project index terms, e.g. "radioactive waste" and "fisheries", and the subprogram types, e.g. "chemistry" and "aeronautics". These terms, especially the project index terms, are partially predefined for the FPs, and need to be selected or added by the people submitting a proposal. Fourth, I reviewed the generic knowledge thesaurus provided in AutoMap and added the entries that seemed relevant in the context of the Funding data to the thesaurus (about 650). Fifth, I automatically deduplicated and manually cleaned all thesaurus entries, e.g. by checking for overly common terms given the domain, and splitting comma separated entries into multiple entries[20]. I re-used and further adjusted the thesaurus cleaning tool built for the Sudan thesauri for this purpose. The resulting Funding master thesaurus contains 4,580 entries. In this thesaurus, all entries are categorized as knowledge, which will result in one-mode network data.

The described thesaurus construction process is a specific example for the more general case of integrating local domain thesauri (in this case derived from salient terms from text data) with standard domain thesauri (in this case FP index terms) and standard generic thesauri (in this case CASOS general knowledge thesaurus). The terminology for these types of thesauri originates from the D2M process description (Carley, Lanham, et al., 2011). Integrating these various types of thesauri is a standard part of the D2M text coding process, and is designed to adapt previously generated thesauri to new domains and datasets.

Completing this process took four work days, and most of the time was spent on a) programming parsers and b) vetting automatically suggested entries for their appropriateness. The outcome of step a) is partially reusable for other projects, while step b) is unlikely to generalize to new domains. Overall, the time costs are a significant decrease from the amount of time needed for building the Sudan master thesaurus (six weeks), and this decrease is mainly due to the one-

[20] The data format for thesauri in AutoMap is .csv. Since entries separated by comma (e.g. rice, rye and wheat) introduce formatting errors into the thesaurus, I put every entry after a comma into a new line.

mode nature of the entries and the fact that fewer previously existing and partially conflicting thesauri had to be integrated.

5.3.2.2 Network Data Extraction from Texts Using the Data to Model Process and Entity Extractor

The same process as described for the Sudan corpus was used to build an auto-generated thesaurus for the Funding data (5.2.2.2). Ultimately, all entries in the Funding thesaurus need to be of type "knowledge", i.e. terms do not need to be classified into meta-network categories once they have been located. In this case, using the boundary detection model would be sufficient to automatically generate a thesaurus. However, since one of my goals here is to evaluate the quality and suitability of the prediction models in application context, I used class model 4 again (meta-network category, specificity, subtype) for creating this thesaurus.

The raw, auto-generated thesaurus had 202,304 entries with a total of 805,035 occurrences. As also observed for the auto-generated Sudan thesaurus, the additional suggestions (N = 27,654) did not seem highly relevant or partially redundant with entries in the regular thesaurus section. Therefore, I again disregarded the additional suggestions. Next, I reviewed the predicted entries in all 44 categories. Table 106 shows these classes along with their accuracy during k-fold cross validation and their size and accuracy (based on my assessment) in the predicted thesaurus (last column in Table 106). The results show that two categories which performed well during K-fold cross validation (resource, money (97.7%) and agent, specific (92.3%)) did not return results of similar accuracy in the application context. This might be due to the fact that these categories have few instances in the funding data such that these classes suffer from sparsity. Moreover, as already observed for the Sudan thesaurus, all categories that I assessed as having medium or bad accuracy in the application context have the specificity value "specific", while "generic" entries are predicted with mainly high accuracy. Table 106 also shows my decision on whether a category was kept in the thesaurus or not. Categories were excluded from further use if their accuracy seemed too low, and/or if their content seemed irrelevant in the context of knowledge networks from funding data. The quantitative impact of all refinement routines described in this section is summarized in Table 107.

Table 106: Application of prediction model to auto-generate thesaurus for Funding corpus

Class labels	**K-fold cross validation**	**Application to Funding data**			
Meta-network category, specificity, subtype	Accuracy rank	Size: Number of examples in thesaurus	Size rank	Assessment of quality	Useful for analysis?

resource, na, money	97.7%	2,792	28	medium	no
location, specific, country	97.0%	15,822	16	good	yes
org-att, specific, nationality	93.8%	20,281	12	good	yes
attribute, na, numerical	93.4%	135,573	1	good	no
time, na, na	93.4%	38,655	6	good	no
event, specific, war	92.6%	26	41	good	yes
agent, specific, na	92.3%	31,146	8	bad	no
organization, specific, gov.	90.8%	29,051	9	good	yes
org-att, specific, political	90.5%	5	44	good	yes
agent, generic, na	90.2%	98,980	3	good	yes
organization, generic, corporate	88.7%	52,534	4	good	yes
location, specific, city	88.1%	12,098	17	good	yes
organization, specific, corporate	87.2%	109,490	2	medium	yes
location, generic, country	87.1%	11,606	18	good	yes
location, specific, state-prov.	85.4%	222	36	good	yes
organization, generic, gov.	81.4%	7,058	20	good	yes
organization, specific, educational	77.8%	3,877	27	good	yes
location, generic, city	77.7%	1,641	31	good	yes
knowledge, specific, law	77.5%	4,356	26	medium	no
organization, generic, educational	72.7%	2,379	30	good	yes
location, specific, other	71.8%	16,423	15	good	yes
resource, generic, product	71.7%	4,808	24	good	yes
event, specific, na	69.0%	626	34	medium	no
location, generic, facility	67.9%	19,410	13	good	yes
organization, specific, other	67.1%	28,081	10	medium	no
attribute, na, age	66.9%	6,062	21	good	no
organization, specific, political	63.8%	31	40	good	yes
resource, na, substance	62.0%	44,124	5	good	yes
organization, generic, other	61.6%	17,982	14	good	yes
org-att, specific, religious	59.6%	10	42	good	yes
location, generic, state-prov.	52.9%	4,942	23	good	yes
resource, na, disease	50.8%	6,042	22	good	yes
knowledge, specific, language	50.0%	735	33	good	yes
location, specific, facility	49.8%	4,646	25	bad	no
knowledge, specific, art	48.5%	26,784	11	medium	no
organization, specific, religious	48.5%	174	37	medium	no
resource, na, plant	48.5%	2,684	29	good	yes
organization, generic, political	48.3%	9	43	good	yes
organization, generic, religious	47.1%	482	35	good	yes
resource, na, animal	40.4%	9,703	19	good	yes
org-att, specific, other	34.4%	96	38	medium	no
task, na, game	29.6%	3	45	good	yes
resource, specific, product	28.0%	33,508	7	bad	no
location, generic, other	18.8%	78	39	good	yes

Next, I applied the same delete list as used for the Sudan thesauri to the Funding thesaurus (hard match on complete entry). Also, I consolidated entries based on their part of speech, subtype, specificity, and meta-network class (Table 107). As already observed for the Sudan data, entries with low frequencies are often long chains of multiple relevant entries. Therefore, I removed all entries with a frequency of one since this seemed a suitable cut-off point.

To further assess the quality of the auto-generated thesaurus, I reviewed all entries with a frequency of 1,000 or more (N = 473). I removed a total of 7 (1.5%) of them as they were overly generic. At this point, the category of "organization, specific, government" still seemed to contain highly generic entries, which I took care of by going through all entries in that category with 1,000 instances or more. Of those unique entries, 7.5% matched in spelling when disregarding capitalization. Since in the next step, all entries were assigned to the same node class (knowledge) or the attribute class, I did not further consolidate entries based on capitalization.

Table 107: Summary of thesaurus cleaning routines and quantitative impact

Routine	Entities		Ratio of raw size	
	Unique	Total	Unique	Total
1. Raw auto-generated thesaurus	202,304	805,035	100%	100%
2. Remove categories with low performance	97,899	497,003	48.39%	61.74%
3. Apply delete list	97,375	466,895	48.13%	58.00%
4. Consolidate entries (in named order) based on part of speech, subtype, specificity, meta-network class	91,480	466,895	45.22%	58.00%
5. Remove entries with frequency of one	17,487	466,895	8.64%	58.00%
6. Correct entries with frequency of 1,000 and more, correct and clean poorly performing categories	17,459	390,344	8.63%	48.49%

After generating one knowledge network for each projects with a text and person per FP, I unionized those networks into one graph and further inspected all nodes with a frequency of 1,000 or more (N = 725). Of those, 80 nodes (11.0%) still seemed overly common. I removed these nodes from the network data directly. I repeated this process again. After this step, I evaluated the network data as not needing further substantial cleaning.

Overall, the process of constructing network data by using the D2M process with the auto-generated thesaurus took about two work days. The reduction of time needed to complete this process from seven days for the auto-generated Sudan thesaurus is due to three reasons:

- Being able to reuse thesaurus post-processing tools that I had built for the Sudan project.

- Repeating the sequence of thesaurus refinement steps that I had identified as being practical, efficient and leading to the intended thesaurus and network data improvements during the Sudan project. However, even though it seems appropriate to reuse these steps, the best parameter setting per step can vary and therefore needs to be tested and adjusted per dataset.
- Generating one-mode networks as opposed to multi-mode networks, where additional time would be needed to verify the classification of entities into node classes and sub-categories, such as specificity values.

In summary, I estimate that comparable time costs of about two days would be necessary to construct and refine a new domain thesaurus with the prediction models under the following conditions:

- The same thesaurus post-processing tools and steps are employed.
- One-mode network data are constructed, regardless of the actual node type.
- The underlying corpus is of comparable size.

5.3.2.3 Network Data Construction from Meta Data

First, for each FP with a person and a text, I created a social network by linking the project coordinator to every collaborator on a given project. Collaborators were not linked to each other in order to avoid overly dense clusters that might not reflect the reality of collaboration on research grants. I made this choice after consulting with faculty who had long-term experience in being the principal investigator on numerous grants. The chosen network formation approach leads to star structures as opposed to complete cliques per project. Stars are networks where nodes link to one central node only. Multiple instances of pairs of collaborating people are reflected in the cumulative edge weight.

Second, I created a knowledge network by linking all unique expressions from the project index terms and subprogram types per project with each other. This results in a clique or complete graph per project. The database fields considered in this step are the same that were used for the building the section of the Funding master thesaurus that uses database entries from CORDIS.

Third, I created an agent by knowledge networks by linking each agent on a project (coordinators and additional collaborators) to each knowledge item per project. All outputs were generated such that they can be loaded as dynamic meta-networks into ORA.

5.3.3 Results

As also observed for the Sudan data, the network size is largely a function of the number of entities considered for network construction (Table 108): since the number of entries in the auto-generated thesaurus (17,459) is larger those in the Funding master thesaurus (4,580) and those considered for meta-data network construction (2,973), the networks produced with the auto-generated thesaurus are the largest. While this finding is intuitive and non-surprising, it needs to be considered when constructing or using thesauri because network size has shown to correlate with network metrics (Anderson, Butts, & Carley, 1999; Faust, 2006; Friedkin, 1981; Marsden, 1990). For example, the larger the network, the lower is the density, and this density value might be independent from the social cohesion of a group, and rather be a result of the number of nodes and possible connections. I conclude that it is important to report the size of thesauri and how the thesauri entries were constructed: the results from the Sudan and Funding data have shown that if thesaurus entries originate from the underlying text data, such as salient text terms, one can expect a higher number of matches and therefore larger networks than when adapting external thesauri to a dataset or domain.

Table 108: Network size per network construction method

FP program	**D2M**		**D2M + EE**		**Meta-data**			
					KK		AA	
	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges
No. of thes. entries	4,580		17,459					
1	1,127	63,832	5,099	235,606	20	23	676	575
2	1,213	90,256	5,414	295,068	91	200	5,547	5,410
3	1,401	118,584	6,079	378,072	295	1,310	14,427	14,251
4	1,623	209,968	8,648	831,452	867	6,447	35,061	34,583
5	1,655	203,350	8,694	754,356	634	9,082	34,541	48,670
6	1,680	179,298	8,146	661,564	1,299	18,888	39,848	43,033
Union	1,945	374,374	12,859	1,949,028	2,923	33,230	117,428	145,898

The results from intersecting the knowledge networks per construction method suggest the following (Table 109): by far, the largest match in nodes *and* edges was observed for the D2M+EE network resembling the D2M network. More specifically, on average, 30.2% of the nodes and 31.2% of the edges contained in D2M are also represented in the D2M+EE network. Even though this effect is non-symmetric, D2M still resembles a comparatively high amount of the links from D2M+EE. Again, one main explanation for the asymmetry might be the size of the respective networks – the D2M+EE networks are about 5.1 times bigger in terms of nodes and 3.8 in terms of links than the D2M network so that the D2M+EE has a larger pool of network constituents that can match the other network.

In contrast to the Sudan application scenario, a larger ratio of nodes contained in the master thesaurus matched the text data (42.5% for Funding versus 11.5% for Sudan). This finding indicates that constructing a domain-specific thesaurus from scratch results in a higher thesaurus coverage rate. For the D2M+EE networks, this ratio is similar for the Sudan data and the Funding data (72.4% and 73.7%); suggesting that the auto-generated thesauri generalize better to new domains.

As already observed for the Sudan project, the meta-data hardly entail any of the links found in the D2M+EE networks (less than 0.7%), but some of the nodes (14.8%) from D2M. An explanation for this finding could be that about 65% of the entities in the master thesaurus (used for D2M) were taken from the same sources (project index terms and subprogram types) as the entities considered in META. None of these sources were used for creating the auto-generated thesaurus (D2M+EE). This rationale would also explain why D2M entails almost 38% of the nodes found in META; the highest resemblance of nodes across all test cases.

In summary, the network size and similarity between thesauri or look-up dictionaries used for network construction seem to be the main factors that determine the overlap of networks. Since the sources for meta-data networks and the auto-generated thesaurus are disjoint pieces of information, these networks share very few constituents. In contrast to that, the master thesauri draws from the sources that are used for identifying nodes for the meta-networks and D2M networks such that overlaps with both types of networks are more likely. However, regardless of this potential "advantage" for the D2M networks, the largest resemblance is still achieved by the D2M+EE networks with respect to the D2M networks; indicating that resemblance can also be identified from the data itself without constructing look-up dictionaries.

Table 109: Overlap between knowledge networks constructed with different methods

FP	Intersection of D2M and D2M+EE				Intersection of D2M and Meta-data (KK)				Intersection of D2M+EE and Meta-data (KK)			
	D2M+EE contained in D2M		D2M contained in D2M+EE		Meta-data contained in D2M		D2M contained in Meta-data		Meta-data contained in D2M+EE		D2M+EE contained in Meta-data	
	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges
1	7.7%	8.9%	35.0%	33.0%	70.0%	17.4%	1.2%	0.0%	35.0%	8.7%	0.1%	0.00%
2	7.9%	9.9%	35.4%	32.2%	45.1%	7.0%	3.4%	0.0%	22.0%	3.5%	0.4%	0.00%
3	7.3%	9.5%	31.5%	30.3%	29.2%	3.7%	6.1%	0.0%	11.2%	2.0%	0.5%	0.01%
4	5.4%	7.2%	28.7%	28.3%	40.4%	5.2%	21.6%	0.2%	11.3%	1.9%	1.1%	0.01%
5	5.5%	7.7%	28.6%	28.6%	29.7%	4.2%	11.4%	0.2%	8.4%	1.6%	0.6%	0.02%
6	5.8%	7.8%	28.0%	28.8%	20.9%	2.4%	16.1%	0.2%	4.5%	0.7%	0.7%	0.02%
Union	3.3%	4.7%	22.0%	24.5%	28.8%	4.0%	43.3%	0.4%	5.7%	1.3%	1.3%	0.02%
Average of years	6.6%	8.5%	31.2%	30.2%	37.7%	6.3%	14.7%	0.1%	15.4%	3.0%	0.6%	0.0%
Rank nodes	5		2		1		4		3		6	
Rank	2		1		3		5		4		6	

links						

In order to test whether any knowledge networks resemble the social networks constructed from the meta-data, I first changed the node type of the social networks to “knowledge”. Otherwise, no matches could be found. The unionized and type-converted network for all FPs (12,859 nodes, 1.95 million links) intersects with the knowledge networks as follows:

- Unionized META: no intersection.
- Unionized D2M: intersect in 1 node and 0 links.
- Unionized D2M+EE: intersect in 144 node and 0 links.

Further looking into the intersection of the social meta-network with D2M+EE suggests that the shared nodes can be references to truly distinct entities that coincidentally overlap in spelling. Examples are “wood” and “benz” in the sense of people (social network) versus entities occurring in the context of a research project (knowledge network). In summary, the outcome from intersecting social networks with knowledge networks suggests that mining the content of text data is not an appropriate strategy for reconstructing social networks. Any agreement between these two types of network might be accidental, such as people’s names coinciding with common nouns.

The results from the key player analysis show that D2M and D2M+EE agree in a few nodes, e.g. “project”, “systems”, “design”, and the shared nodes even rank similarly with respect to centrality metrics (Table 110). The meta-data knowledge networks do not overlap in key entities with the text-based knowledge networks. Even though all three types of networks contain very domain-specific terms, the most prominent entities in D2M and D2M+EE are rather generic entities from the research domain, while the key entities from META refer to more specific research areas. This difference might be explained by the data sources: the meta-data entities originate from key words, which are highly concise summaries of the content of a project description, while the text bodies of these descriptions explain the projects in more detail. Taking this last point together with the low intersection rate of meta-data networks with text-based networks (at least on the link level), it seems recommendable to combine both types of networks to cover both, the common terms of a corpus and domain as well as specific, higher-level aggregates of the corpus’ content. Since D2M+EE resembles about a third of D2M and lead to similar types of key entities as D2M, and D2M already partially overlap with the META, it might suffice to combine just the D2M+EE networks plus the meta-data networks for this purpose.

Table 110: Key entities per network construction method (networks unionized for all FPs*

Degree Centrality				Betweenness Centrality			
Key entity	D2M	D2M+EE	Meta-data	Key entity	D2M	D2M+EE	Meta-data
project	1.3	1.0		project	1.3	1.0	
development	3.0			development	2.7		
european	4.0			research	3.3		
system	4.0			european	4.7		
research	4.3			europe	5.0		
develop	5.7			systems	6.0	2.0	
systems	6.3	2.7		developed	7.7		
information	8.7			develop	8.0		
data	9.3			order	9.0		
design	9.7	9.3		system	9.7		
process	11.7			application	11.3		
developed	12.0			information	11.7		
results	12.3			study	12.7	10.0	
analysis	12.7	5.3		data	13.3		
model	15.0	8.0		results	13.7		
europe		3.3		design		3.3	
study		3.7		analysis		3.7	
countries		7.7		methods		5.7	
studies		8.3		applications		7.7	
applications		8.7		tools		7.7	
field		10.0		techniques		8.0	
methods		12.3		software		10.0	
potential		12.3		field		10.3	
level		12.7		materials		11.0	
techniques		14.7		models		12.0	
				model		13.3	
				studies		14.3	
scientific_research			1.3	environmental_protection			2.7
social_aspects			3.0	policies			5.0
industrial_manufacture			5.3	social_aspects			6.3
information_processing			5.7	safety			6.7
information_systems			5.7	training			6.7
environmental_protection			6.3	renewable_sources_of_energy			7.0
training			7.0	standards			7.3
education			7.3	biotechnology			8.0
electronics			9.0	scientific_research			8.3
microelectronics			9.0	industrial_manufacture			8.7
safety			9.3	technology_transfer			8.7
renewable_sources_of_energy			10.7	information_processing			9.3
other_energy_topics			11.7	waste_management			10.7
materials_technology			12.0	information_systems			11.3
waste_management			14.0	telecommunications			13.3

* Top 15 key entities considered. Values are ranks (1 = highest) averaged over FPs 4 to 6 per metrics. Data from FP1 to 3 not considered in this table because these networks are so small that less than 15 key entities were found.

5.4 Application Context III: Enron Corpus

From its formation in 1985 until mid 2001, the Enron Corporation ("Enron") was a highly and internationally successful trader and broker of energy, commodities and stock options. A combination of unethical to illegal business practices, such as booking losses to "special purpose entities" that did not appear on the companies' public financial reports, and a corporate culture of making risky investment allegedly led to the abrupt fall of Enron (Fox, 2003; Fusaro & Miller, 2002; Powers, Troubh, & Winokur, 2002) (for a more detailed description of the Enron story see also (Diesner et al., 2005)). In December 2001, the company filed for Chapter 11 bankruptcy, which was followed by broad public outcry and uproar among Enron's stakeholders. Both the Federal Energy Regulation Commission (FERC) and the US Security and Exchange Commission (SEC) started investigations into Enron. A by-product of these investigations was the release of the Enron dataset (described below). People have used been using the Enron data to answer substantive question about business networks such as:

- How is covert information disseminated in an organization, and how does the flow of covert information relate to the network structure of an organization? (Aven, 2010)
- How do the properties and structure of communication networks change during an organizational crisis? (Diesner & Carley, 2005a)
- How does the formal structure of an organizational relate to the information structure of the communication network, and how does this relationship change during a crisis? (Diesner et al., 2005)

5.4.1 Data[21]

The Enron email dataset was originally released online by the FERC in May 2002. FERC made the data available in order to allow the public to understand why they had started investigations into Enron. It is crucial to stress the fact that this dataset contains data from many individuals who were not involved in any of the actions that were subject to the Enron investigation.

Each email contains three sources for network data:

- Explicit relational data provided in the email headers, i.e. the email addresses of the senders and receiver(s).

[21] The description of the Enron dataset is based (Diesner et al., 2005).

- Text bodies, which may contain explicit and implicit descriptions of relationships between socio-technical entities.
- Additional meta-data, such as time stamps and folder names.

FERC collected a total of 619,449 emails from 158 Enron employees, mainly from senior managers. The original version of the dataset had a variety of integrity problems. Next, Leslie Kaelbing from MIT purchased the data. The data was then acquired by researchers from SRI, notably Melinda Gervasio, who fixed many of the integrity problems and released their version of the dataset online. In March 2004, William Cohen from CMU put the data online for research purposes. Cohen's version of the dataset contains 517,431 distinct emails from 151 unique users. These emails are organized in 150 user folders with a little less than 4,700 subfolders. Some messages were deleted in response to requests from affected employees. Invalid email addresses for which a recipient was specified were converted to addresses of the form "user@enron.com", and to "no_address@enron.com" where no recipient was specified. Further consistency checks by Andres Corrada-Emmanuel from the University of Massachusetts via applying check-sums (MD5) to email bodies revealed that the corpus actually contained 250,484 unique emails from 149 people.

We started off building the CASOS Enron database by using the version provided by Jitesh Shetty and Jafar Adibi from ISI. The ISI researchers had refined and normalized the dataset by dropping blank, duplicated and junk emails, and emails that had been returned by the system due to transmission errors. The resulting corpus consists of 252,759 emails organized in 3,000 user defined folders from 151 distinct people. The ISI group put the Enron data in a MySQL database which contains four tables; one for *employees*, *messages*, *recipients* and *reference information*. We chose this version of the dataset for our work because the normalization processes that were done to it seemed appropriate to us and were well documented and the data structure met our needs. I refer to this version of the Enron email dataset as the "CASOS Enron dataset".

This dataset also involved a co-reference resolution challenge: the entities or nodes represent email addresses, not people. This is troublesome for cases in which people use more than one email address, such that unique individuals would occur as multiple nodes in the network. We have corrected for this issue by mapping e-mail addresses to individuals based on information about Enron employees as provided in publically available data sources. These external data sources contain information about the location of the Enron branches that people worked in, as well as their job titles. For a full description of the preparation of the CASOS Enron dataset see (Diesner et al., 2005). In summary, we were able to map 1,234 email addresses to 557 distinct individuals for who we also know their real name. In these refined data, the number of email addresses per person ranges from 1 to 17, the average number of emails per person is 2.2, and the standard deviation for this number is 1.9. The number of emails for which both a sender and at least one receiver can be mapped to a unique and disambiguated individual is 52,866 (21.1% of the number of unique emails identified by Corrada-Emmanuel). We equally consider entries in the *to*, *cc*, and *bcc* fields as receivers. This version of the CASOS Enron dataset is used herein for analysis.

For the previous two application scenarios, the time slicing of the corpora was done based on calendar years (Sudan corpus) and funding periods (Funding corpus). The first approach could also be used for Enron. However, since the Enron data offer a rare glimpse into a real-world, organizational crisis, I decided to construct time slices around critical periods in Enron's history, even though no empirical questions about the Enron crisis are addressed herein. The Enron crisis started in August 2001, when Jeffrey Skilling suddenly resigned as CEO, and Kenneth Lay took over this position again. In the same month, Sherron Watkins, one of Enron's vice presidents, wrote a whistle-blower letter to Lay. The crisis then took off in October 2001, when Enron began to publically report its humongous losses. The stock market reacted with a sharp drop in the price for Enron shares; which ultimately led to the company's insolvency. Based on this timeline, I created three time periods that are used herein:

- May to June 2001: 6,091 emails. This period can be considered as a control case. During this period, Enron's fall was not yet in sight.
- August – September 2001: 3,711 emails. The period in which the Enron crisis emerged.
- October – December 2001: 11,042 emails. The period of Enron's downfall.

Taken together, the emails in these three time periods account for 41.0% of all emails in the CASOS Enron dataset.

5.4.2 Network Data Construction Methods

The same methods for network data construction as used for the Sudan and Funding corpus were also used for the Enron corpus where possible.

5.4.2.1 Network Data Extraction from Texts Using the Data to Model Process

I started to create the Enron master thesaurus by reusing multiple local domain thesauri that we had previously built for the CASOS Enron data by using the D2M process. For this D2M process, we had employed an earlier entity extractor that I had also built for AutoMap by using conditional random fields-based machine learning techniques (Diesner & Carley, 2008a). After combining the various local domain thesauri, I added standard domain thesauri for Enron which contain the names of people. These thesauri were generated from the explicit meta-data in the email headers that specify the senders and receivers of emails. Finally, I enhanced the Enron master thesaurus with entries from the standard generic thesauri that are provided in AutoMap: I reviewed the entries in the standard thesauri for agents, organizations, events, tasks, knowledge, locations, roles (generic agents) and time one by one, and added the entries that I considered as relevant to the master thesaurus. Like the Sudan thesauri, this thesaurus is multi-modal, i.e. nodes are placed into different meta-network categories. In fact, some entries from the local domain thesauri for Enron also occurred in the standard generic thesauri, such that these thesauri had some overlaps, which I removed.

After generating and inspecting the D2M network data, I identified a few more nodes that appeared as key players, but for which the overlap in case-insensitive spelling with other, more common terms had contributed to the high frequency and prominent network position of these nodes. An example is "price", which is the last name of a former Enron employee, but the term is more often used as a common noun, e.g. to refer to the price of shares. I removed these nodes from the master thesaurus and regenerated the network data. The final Enron master thesaurus contained 6,963 entries.

Completing the construction of the Enron master thesaurus took two work days. As already observed for the Funding master thesaurus, reusing and adapting existing thesauri significantly cuts the time costs for thesaurus construction.

5.4.2.2 Network Data Extraction from Texts Using the Data to Model Process and Entity Extractor

Class model 4 was used again to produce the auto-generated Enron thesaurus. The raw thesaurus contained 144,204 entries with a total of 633,597 instances. Like in the previous applications scenarios, I disregarded the additional suggestions (N=9,228) for the same reasons as outlined before. Again, I reviewed each category separately. Table 111 shows the outcome of this process and also specifies which categories were not further considered due to low performance during application.

Table 111: Application of prediction model to auto-generate thesaurus for Enron corpus

Class labels	**K-fold cross validation**	**Application to Funding data**			
Meta-network category, specificity, subtype	Accuracy rank	Size: Number of examples in thesaurus	Size rank	Assessment of quality	Used for analysis?
resource, na, money	97.7%	19,228	9	good	yes
location, specific, country	97.0%	2,528	21	good	yes
org-att, specific, nationality	93.8%	920	26	good	yes
attribute, na, numerical	93.4%	98,886	2	good	yes
time, na, na	93.4%	76,008	3	good	yes

event, specific, war	92.6%	17	42	good	yes
agent, specific, na	92.3%	60,220	4	medium	yes*
organization, specific, gov.	90.8%	518	29	good	yes
org-att, specific, political	90.5%	98	39	good	yes
agent, generic, na	90.2%	38,565	6	good	yes
organization, generic, corporate	88.7%	23,098	8	good	yes
location, specific, city	88.1%	11,966	11	good	yes
organization, specific, corporate	87.2%	2,167	22	good	yes
location, generic, country	87.1%	1,083	25	medium	no**
location, specific, state-prov.	85.4%	1,422	24	good	yes
organization, generic, gov.	81.4%	4,214	18	good	yes
organization, specific, educational	77.8%	10,705	12	good	yes
location, generic, city	77.7%	479	31	good	yes
knowledge, specific, law	77.5%	8,964	14	good	yes
organization, generic, educational	72.7%	545	27	good	yes
location, specific, other	71.8%	5,395	16	good	yes
resource, generic, product	71.7%	437	34	good	yes
event, specific, na	69.0%	486	30	bad	no
location, generic, facility	67.9%	4,077	19	good	yes
organization, specific, other	67.1%	9,979	13	medium	no**
attribute, na, age	66.9%	4,793	17	good	yes
organization, specific, political	63.8%	450	33	good	yes
resource, na, substance	62.0%	1,479	23	good	yes
organization, generic, other	61.6%	6,043	15	good	yes
org-att, specific, religious	59.6%	10	44	good	yes
location, generic, state-prov.	52.9%	3,835	20	good	yes
resource, na, disease	50.8%	531	28	bad	no
knowledge, specific, language	50.0%	61	41	good	yes
location, specific, facility	49.8%	16,956	10	medium	yes*
knowledge, specific, art	48.5%	25,871	7	bad	no
organization, specific, religious	48.5%	155	35	bad	no**
resource, na, plant	48.5%	100	38	good	yes
organization, generic, political	48.3%	148	36	good	yes
organization, generic, religious	47.1%	146,747	1	bad	no**
resource, na, animal	40.4%	470	32	medium	no**
org-att, specific, other	34.4%	16	43	good	yes
task, na, game	29.6%	82	40	good	yes
resource, specific, product	28.0%	43,734	5	bad	no
location, generic, other	18.8%	111	37	good	yes

* entries with frequency of 50 and more reviewed and corrected if needed, all entries maintained

** entries with frequency of 50 and more reviewed and corrected if needed, all other entries deleted

Next, I refined the auto-generated thesaurus as summarized in Table 112. Then, I used the refined thesaurus to extract meta-networks from the email bodies by employing the D2M process. I further refined the thesaurus by reviewing all nodes in the resulting networks with a frequency of at least 100 (N=1,167). Based on this review, I deleted overly common entries from the thesaurus and modified category assignments where needed. Regenerating and inspecting the nodes suggested that the thesaurus and network data were sufficiently clean at this point, particularly for highly frequent nodes. Overall, post-processing the auto-generated Enron thesaurus took about two work days, which is comparable to the time costs for building a master thesaurus from existing sources.

Table 112: Summary of thesaurus cleaning routines and quantitative impact

Routine	Entities		Ratio of raw size	
	Unique	Total	Unique	Total
1. Raw auto-generated thesaurus	144,204	633,597	100%	100%
2. Remove categories with low performance	66,330	386,737	46.0%	61.0%
3. Apply delete list	66,068	360,896	45.8%	57.0%
4. Consolidate entries (in named order) based on part of speech, subtype, specificity, meta-network class, spelling regardless of capitalization	60,373	360,896	41.9%	57.0%
5. Remove entries with frequency less than five	8,549	275,952	5.9%	43.6%
6. Correct entries with frequency of 100 and more	8,546	275,497	5.9%	43.5%
7. Correct entries after reviewing nodes with frequency of 100 and more in unionized graph (N = 1,167), re-deduplicate nodes	8,255	272,647	5.7%	43.0%

Table 113 shows the frequency distribution of node classes in the final auto-generated thesaurus. As also observed for the Sudan data, overall, generic social agents (individuals and groups) occur more often in the text data than specific agents. This finding further supports the importance of considering unnamed entities for socio-technical network analysis in addition to the traditional focus on specific entities.

Table 113: Frequency distribution of entities classes in thesaurus*

Class	Ratio in full thesaurus, unique	Ratio in full thesaurus, total	Average number of repetitions per unique entity
agent, specific	26.9%	10.9%	13.4
attribute	24.6%	28.2%	37.9
time	16.8%	19.7%	38.7
resource	7.5%	3.7%	16.3
agent, generic	7.0%	12.8%	60.7
location, specific	6.7%	6.2%	31.0
organization, specific	3.5%	4.1%	38.5
knowledge, specific	2.8%	1.1%	12.5
organization, generic	2.7%	11.4%	137.1

location, generic	0.8%	1.6%	66.2
knowledge	0.4%	0.1%	10.6
resource, generic	0.2%	0.2%	22.3
task	0.1%	0.0%	14.8
Total	100.0%	100.0%	33.0

* values over 10% underlined

Reviewing the auto-generated Enron thesaurus and respective D2M+EE network data at different stages of refining the thesaurus, I made the following observations:

First, I had hypothesized that since the Enron data are from a different time period, domain, and writing style than the data used for training the prediction models, the prediction accuracy would be lowest for this application scenario. The results do not support this hypothesis: based on my qualitative reviews presented in this chapter, the prediction accuracy was about the same across all three corpora, with the same classes being problematic throughout.

Second, the errors made by the prediction models are similar across all three applications:

- A most commonly observed type of error was the assignment of terms that typically occur in lower case to classes of specific agents or specific organizations for cases in which these terms were capitalized. This happens if the impacted terms appear at the beginning of a sentence or when all letters are in upper cases, such as for acronyms (in Sudan and Funding corpora) and "yelling" in emails (Enron).
- Erroneous cases with a low frequency per class (less than ten, especially one to five) often involve chains of multiple entities (Sudan, Funding) or of relevant entities in conjunction with highly frequent, domain specific terms, such as "subject" and "Forward" (Enron).
- Specific entities are predicted with a lower accuracy than a) generic entities and b) entities to which the specificity distinction does not apply.
- Categories performing low during formal model testing are more likely to also perform low when applying the models to new and unseen data; with two exceptions:
    - Categories that performed very well during formal model assessment might return poor results during application, especially for specific agents.
    - Categories that performed low during formal model assessment might return good results during application.

5.4.2.3 Network Data Construction from Meta Data

Similar to the procedure used for the Funding data, I built the meta-networks from the information explicitly given in the email headers: I used the information about senders and receivers to generate directed social network. This information was also used as standard domain thesauri for the Enron master-thesaurus (used for D2M networks). The weight of a link is the number of emails exchanged between the involved agents. Any type of receiver (to, cc, bcc) is equally considered as an email recipient. Even though these social networks might be incomplete since not all of Enron's emails are present in the dataset, they can be considered as a type of ground truth data.

5.4.3 Results

Table 114: Network size per network construction method

Data	D2M		D2M+EE		Meta-data		Number of emails
	Nodes	Edges	Nodes	Edges	Nodes	Edges	
No. of thes. entries	6,963		8,255				
Pre-crisis	1,504	27,618	3,506	54,846	448	3,092	6,901
Emergence of crisis	1,547	21,071	3,149	43,452	433	2,295	3,711
Crisis	1,665	31,624	3,989	71,068	435	4,721	11,042
Union graph	1,940	55,956	4,794	132,064	513	7,365	21,653

The auto-generated thesaurus contains 1.2 more entries than the master thesaurus, but leads to the retrieval of 2.3 more nodes and 2.1 more edges (Table 114). Also, 58.1% of the entities in the auto-generated show up in the D2M+EE networks, while 27.9% of the entries from the master thesaurus appear in the D2M networks. This indicates again that the auto-generated thesaurus is not only more efficient to build, but also more effective to use than the master thesaurus.

Another crucial finding is that the text-based networks contain 8.4 (D2M+EE) and 3.6 (D2M) more nodes than the meta-data networks. This effect is not necessarily evident from the density values of the networks (Table 115), which are almost identical for the meta-data networks and the text-based networks. Nonetheless, this finding indicates that the windowing technique for link formations applied to network data generates more dense networks than the social networks from the email headers, which can be considered as ground truth data.

Table 115: Network density per network construction method

Data	D2M	D2M+EE	Meta-data
Pre-crisis	0.02	0.01	0.02
Emergence of crisis	0.01	0.01	0.01
Crisis	0.02	0.01	0.03
Union graph	0.02	0.01	0.02

In order to analyze the structural overlap of the meta-data networks with the text-based networks, I extracted only the connections among specific agents as they resemble the same type of nodes as the entities in the meta-data. Applying this constraint, the intersection between the meta-data networks (proxy for ground truth) and the text-based networks is particularly high on the node level for the D2M networks resembling the meta-data networks (86.8%), and moderately high for the vice versa case (54.9%) (Table 116). This result is intuitive because all of the entities contained in the meta-data network were also added as entries to the master thesaurus and most of the specific agents in the master thesaurus originate from this set of entities. Since the list of email senders and receivers was not added to the auto-generated thesaurus, the mutual resemblance of the meta-data networks and the D2M networks is minimal.

Table 116: Overlap between social networks (agents, specific only) constructed with different methods

Data	Intersection of D2M and Meta-data				Intersection of D2M+EE and Meta-data			
	D2M contained in Meta-data		Meta-data contained in D2M		D2M+EE contained in Meta-data		Meta-data contained in D2M+EE	
	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges
Pre-crisis	60.1%	9.6%	88.4%	19.6%	6.7%	0.06%	2.4%	0.02%
Emergence of crisis	53.6%	7.0%	83.6%	14.4%	6.0%	0.09%	2.3%	0.02%
Crisis	51.1%	10.6%	88.5%	17.5%	6.7%	0.06%	1.9%	0.02%
Union graph	57.7%	12.0%	94.0%	22.7%	6.8%	0.15%	1.9%	0.04%
Average of years	54.9%	9.1%	86.8%	17.2%	6.5%	0.1%	2.2%	0.0%

Comparing the text-based networks of specific agents shows that even though no shared entries were explicitly added to both thesauri, both networks still pick up on a small amount of common agents (left-hand side section in Table 117). In order to test for the overall structural agreement between the text-based networks, I also considered all node classes for comparison, including but not confined to specific agents (right-hand side section in Table 117). This comparison shows that D2M+EE resembles D2M more than vice versa to almost the same amount as D2M+EE is larger in nodes and edges than D2M. This finding further confirms the prior observation that structural overlap correlates with network size.

Table 117: Overlap between networks constructed with different methods

Data	Intersection of D2M and D2M+EE Agent, specific network				Intersection of D2M and D2M+EE Entire meta network			
	D2M contained in Meta-data		Meta-data contained in D2M		D2M contained in D2M+EE		D2M+EE contained in D2M	
	Nodes	Edges	Nodes	Edges	Nodes	Edges	Nodes	Edges
Pre-crisis	10.2%	1.5%	5.3%	0.7%	18.9%	4.4%	8.1%	2.2%
Emergence of crisis	9.6%	1.1%	5.7%	0.5%	18.4%	3.8%	9.0%	1.8%
Crisis	9.6%	1.7%	4.6%	0.8%	18.5%	4.0%	7.7%	1.8%

Union graph	9.0%	1.6%	4.0%	0.7%	16.8%	4.3%	6.8%	1.8%

For Enron, key player analysis was conducted on the level of specific agents, because this is the only class of nodes that is available in all three types of networks. The meta-data networks and D2M networks share almost the same list of thesaurus entries or entities considered for network construction, and most of the key players in D2M originate from this list of agents (77.5% on average, those with first and last name). However, there are hardly any overlaps in key players between META and the text-based networks (Table 118, Table 119).

Taking the findings from the structural agreement and overlap in key players together, it seems that even though some networks per construction methods have significant intersections, they lead to different suggestion about who the main agents in a network are.

For the Sudan application scenario it had been shown that the key players from text-based networks are often first names of specific agents. While these nodes get assigned to the correct node class, they often cannot be associated with specific individuals who have a first *and* last name. This issue is even more likely to occur in the Enron data, since in the US-American business culture, people often address and refer to others by their first name and also sign emails with their first name. The results shown in Table 119 confirm this assumption for the D2M+EE networks, and to a lesser degree also for the D2M networks. In fact, most occurrences of specific agents with a first *and* last name are likely to originate from email headers that occur in email bodies due to the forwarding of emails and to a lesser degree also from email signatures, which are not very common among the internal emails in Enron. Therefore, the results suggest that with the master thesaurus (D2M), it is more likely to retrieve names from meta-data within the text bodies, while with the auto-generated thesaurus (D2M+EE), instances of first names only, which are more likely to occur in the actual content of an emails, are more often identified as key agents. D2M might not even pick up on these names if they do not occur in the master thesaurus. As described for the Sudan thesaurus, mapping these agents to a first *and* last name might be infeasible since multiple people might share the same first name.

Table 118: Key agents per network construction method I

Degree Centrality				**Betweenness Centrality**			
Key Entity	D2M	D2M+EE	Meta	Key Entity	D2M	D2M+EE	Meta
lloyd_will	2.0			lloyd_will	1.0		6.3
rebecca_mark	2.3			rebecca_mark	2.0		
jeff	2.7	1.3		jeff	4.0	1.7	
jeff_dasovich	3.7		2.3	dorland_chris	4.0		
thomas_paul_d	5.0			susan_scott	6.3		
kean_steven_j	6.0			dave	6.7		

steven_kean	7.7			eric	6.7		
paul_kaufman	8.0			mathew_frank	7.7		
susan_scott	8.7			kean_steven_j	8.3		
dorland_chris	9.0			thomas_paul_d	8.3		
james		2.7		john		3.7	
john		3.3		jim		5.0	
richard		4.3		mike		5.0	
dasovich		5.0		richard		5.0	
steffes		5.0		james		6.3	
steve		8.0		kim		6.3	
susan		8.0		steve		6.7	
mike		8.7		jones		7.0	
shapiro		8.7		chris		8.3	
susan_mara			4.3	louise_kitchen			2.7
james_steffes			4.7	john_lavorato			3.0
louise_kitchen			5.0	timothy_belden			4.7
mike_grigsby			5.7	kevin_presto			5.0
mary_cook			6.0	mark_haedicke			5.3
richard_shapiro			6.3	tom_may			6.7
liz_taylor			6.7	christi_nicolay			7.0
iris_mack			7.0	kay_mann			7.0
john_lavorato			7.0	mark_taylor			7.3

Table 119: Key agents per network construction method II

Eigenvector Centrality				**Clique Count**			
Key Entity	D2M	D2M+EE	Meta	Key Entity	D2M	D2M+EE	Meta
jeff_dasovich	1.3		1.7	rebecca_mark	1.3		
jeff	1.7	3.0		lloyd_will	1.7		6.7
thomas_paul_d	3.3			jeff	3.0	3.3	
paul_kaufman	4.3		6.0	susan_scott	5.0		
lloyd_will	4.7			thomas_paul_d	5.3		
richard_shapiro	7.0		3.7	dorland_chris	5.7		
jeff_richter	7.0			elizabeth	8.0		
rebecca_mark	7.7			kean_steven_j	8.0		
alan_comnes	8.7			mathew_frank	8.3		
alan	9.3			dave	8.7		
james		3.0		john		1.0	
dasovich		4.0		james		3.0	
steffes		4.0		robert		4.3	
richard		4.7		steve		4.7	
shapiro		5.0		richard		5.0	
mara		6.0		mike		6.7	

susan	6.3		tom	8.7	
linda	9.3		jim	9.0	
john	9.7		chris	9.3	
susan_mara		3.3	louise_kitchen		3.3
james_steffes		4.0	john_lavorato		4.0
mary_cook		6.3	kevin_presto		4.3
steven_kean		6.3	timothy_belden		4.3
marie_heard		7.3	christi_nicolay		6.0
harry_kingerski		8.0	mark_haedicke		6.0
mark_palmer		8.3	steven_kean		6.3
			don_baughman		7.0
			liz_taylor		7.0

5.5 Conclusions

The application scenarios presented in this chapter are representative for situations in which there is a need for distilling information about relevant entities and their relations from text data, and where the definition of what is "relevant" varies depending on the research question and context. What is generally needed in such situations is the transformation of text data into concise, accurate and reliable reductions and abstractions of the original material, in this case from text data into network data. The results from this chapter suggest the following answers to my research questions:

1. *How do certain prediction models evaluated with k-fold cross validation perform in real-world application scenarios?*

The assessment of auto-generated thesauri across three application scenarios that differ among each other and from the learning data in domain, writing style and publication date of the text as well as of the network data constructed by using these thesauri as part of the relation extraction process lead to the following conclusions:

1. For the majority of the entity classes supported by the prediction models (N = 44 at most), instances are predicted with an accuracy that is high enough for being employable in practical applications to new datasets and domains.
2. In contrast to my initial hypothesis, no meaningful differences in prediction accuracy were observed for publication times, genres and writing styles that differ from those of the learning data.
3. The auto-generated thesauri generalize better to new datasets and domains than the master thesauri, which are built in a more manual fashion.

4. The auto-generated thesauri are more efficient (in terms of construction time) and effective (in terms of picking up matches from the text data) than the master thesauri.
5. As observed in chapter 3 for formal prediction model assessment, the prediction accuracy of classes seems to be independent of the number of instances per class.
6. The auto-generated thesauri feature limitations with respect to prediction accuracy. Therefore, it seems recommendable to verify and if needed correct the auto-generated thesauri. In this chapter, heuristics, methods, and tools were developed to aid this process.
7. Classes that perform low during formal model assessment are more likely to show low performance in the application scenarios as well. However, classes with high accuracy during formal model assessment can return poor results during application and vice versa. The implications of this finding is that is seems recommendable to:
   - Verify the performance of each class in the application context.
   - If the verification of all classes is not feasible, e.g. because it is too time consuming, disregard the classes that perform poorly across all three application scenarios (named below).
8. Several classes show poor performance across all three application scenarios. Since these scenarios involved data from different times, domains and writing styles, the poor performance of these classes might generalize to other datasets:
   - agent, specific
   - organization, specific, corporate
   - event, specific
   - location, specific, facility
   - knowledge, specific, art
   - resource, specific, product
9. Especially for social agents, specific entities are predicted with a lower accuracy than a) generic entities and b) entities without a specificity value. This might be due to data sparseness, i.e. a lower number of specific entities than generic ones of a certain node class in the text data. This assumption is supported by the findings from this chapter.
10. Prediction accuracy per entity drops with cumulative decreasing frequency, i.e. the number of times that an entity is observed in a particular class and – if applicable – further sub-categories, i.e. specificity and subtype.
11. Two main types of errors were observed for the auto-generated thesauri across all three application scenarios:
    - Terms that typically occur in lower case get assigned to the wrong category (mainly specific agents and organizations) when they occur in capitalized form. This might be due to data sparseness, and mainly happens if these terms occur at

the beginning of sentences or when all letters of a term are capitalized, e.g. for acronyms and “yelling” in emails. These cases can be removed from the thesauri by working through the following process: compare the spelling and part of speech of any two entities, output the cases that differ only in capitalization, and make a decision about these cases by manually vetting them or automatically preferring those with a higher frequency count (those counts are included in the auto-generated thesauri).

    - Terms with a low cumulative frequency (less than ten, especially one to five) often involve chains of multiple entities of relevant entities in conjunction with highly frequent, domain specific terms. These entries can be removed from the thesauri by disregarding suggestions with low frequencies. Again, this decision should be based on screening the thesaurus and identifying a suitable cut-off value.

12. Entries in the agent generic and organization generic classes tend to overlap for the case of references to groups, such as “students” or “workers”. In the CASOS standard thesauri, such entries also occur in either thesaurus category. For practical applications, it seems justifiable and efficient to merge these two classes.

2. *How do the network data and respective analysis results obtained by conducting relation extraction including using the entity extractor developed in this thesis compare to alternative methods for constructing network data from the same corpora?*

The comparisons of network data generated with different methods on the structural level and with respect to key entities lead to the following conclusions:

1. Ground truth data constructed by subject matter experts are hardly resembled by any automated methods that analyze text bodies, and are not resembled by exploiting existing meta-data from text corpora. Trying to reconstruct social network data from the content of text body will lead to largely incomplete networks.
2. Networks extracted from text bodies by using auto-generated thesauri (D2M+EE networks) resemble networks generated with master thesauri (D2M networks) more strongly in terms of nodes and edges than vice versa.
3. D2M networks resemble meta-data networks (META) more closely than D2M+EE networks. The main explanation for this finding is that in this study, master thesauri were

enhanced with information from the same sources that were used for defining the nodes in META. At the same time, auto-generated thesauri and meta-data networks are built from disjoint pieces of information, namely text bodies and meta-data on the texts.

4. Agreements in structure and key entities are mainly impacted by two factors:
    - Network size: the larger a network, the higher is the chance that the network resembles parts of network data constructed with other methods. This finding is relevant as it has been shown that network metrics can correlate with network size (Anderson et al., 1999; Faust, 2006; Friedkin, 1981; Marsden, 1990). Consequently, observed differences in these metrics across networks constructed with different methods might be independent of differences in the underlying network, but rather be a consequence of the network construction method, in the case of this study especially the link formation methods.
    - Overlaps in thesaurus content: similarity in entities considered in the thesauri (text-based networks) or for network construction (meta-data networks) strongly impacts the agreement in structure and key players.
5. Structural agreements are always considerably higher on the node level than on the edge level. However, this finding is heavily impacted by the link formation methods used in this chapter. The limitations of this approach were measured and summarized in chapter 2.
6. Meta-data networks are less likely than text-based networks to suffer from co-reference resolution issues. This is mainly because somebody or some algorithm has already solved this issue to some degree. In contrast to the meta-data networks, both types of text based networks (D2M+EE, D2M) tend to retrieve single first names as prominent key entities, and it can be hard to impossible to map these entities to unique people for whom at least a first name and a last name are known.
7. For social networks (agents and organizations) constructed from news wire data, meta-data networks are more suited for providing an overview on major international key entities and their relations. Social networks extracted from text bodies are more appropriate for gaining a localized view on geo-political entities and retrieving information about their culture.
8. Meta-data networks contain more specific entities (in a qualitative, not quantitative sense) than the text-based networks. For the case of knowledge networks, meta-data networks entail more informative key entities than text-based networks, while text-based networks identify many common place terms of a given domain as key entities.
9. Overall, it seems recommendable to combine meta-data networks with text-based networks to cover both, common yet highly salient terms in a domain with more specific,

domain dependent information. For this purpose, it might suffice to combine the networks built with auto-generated thesauri (D2M+EE) with the meta-data networks plus any information from subject matter experts if available. This recommendation is based on the following reasons:

- The D2M+EE networks resemble the D2M networks better than vice versa.
- The D2M+EE networks lead to similar types of key entities than the D2M networks.
- The D2M networks already partially overlap with the meta-networks.

5.6 Limitations and Future Work

The knowledge gained from this chapter is limited by the datasets that I had collected, prepared and used herein, and the methodological choices I made. I discuss both point below and suggest solutions for practical applications with the given methods and technologies as well as ideas for improving these methods and technologies in future work.

5.6.1 Data Level

Even though the Sudan corpus was collected through LexisNexis from a variety of sources, most of the texts are from newspapers and news magazines published in English. The biases that are contained in these sources are carried over to the extracted network data. Especially the analysis of meta-data had shown that one of these biases is a focus on high-profile politicians from the Western world.

The CORDIS database might be incomplete, i.e. some funded project might be missing. There is no way for us to validate the completeness of the provided information. Also, the database is incomplete for some of the listed projects. Moreover, the CORDIS database does not list rejected proposals, and no public source might provide this information. Also, the co-reference procedures that I applied to the individuals in the data leave room for improvements: errors such as typos could be further eliminated by employing edit-distance algorithms. Also, detecting variations in names due to name changes, e.g. when women adopt their husband's last name, would require further careful checking of institutional affiliations and addresses.

The Enron data are also likely to be incomplete as only the email archives from 158 people were collected, and people might not have stored all of their emails in these archives. Similar to the limitations pointed out for the cleaning of the Funding data, the data cleaning process might be incomplete: people with identically spelled names and email addressed might have been aggregated, people for who we could not map a real name to one or more email addresses were disregarded from analysis, and people included in the analysis might have used additional email addresses that we were not able to associate with them. However, the advantage with the CASOS Enron email dataset is that nodes represent individual people as opposed to email addresses. This might entail the risk of conflating various "personas" or roles that people play when using different email addresses, such as one for professional and one for private matters.

5.6.2 Methods Level

Various methodological limitations apply to the conclusions drawn from this chapter:

1. Automated text coding: Even though automated text coding speeds up computer-assisted and even more so manual text coding, it involves various weaknesses: entity extraction tools are more likely than humans to retrieve duplicates and near duplicates (Bond et al., 2003). This was also observed in the application contexts. On the other hand, machine coding offers perfect intercoder-reliability (at least for non-probabilistic methods) and excludes accuracy losses due to fatigue and coding biases due to individual contextualization or interpretation of the data (Schrodt, 2001).

2. Impact of human decisions and need for subject matter expertise: Even though many of the text coding and network analysis routines used in this chapter are largely supported by software tools, there are still numerous manual and computer-assisted steps involved. These steps are not only time consuming, but also require human decision making processes. It was shown that such processes imply the risk of errors and reliability issues and biases (chapter 2), and require substantial subject matter expertise. In this chapter, a single person (me) made these decisions and tried to acquire subject matter expertise on domains and datasets as needed. This might be representative for real-world text coding projects. However, the following strategies were used to mitigate the named risks: all decisions were made in close coordination with my advisor, according to the norms and rules established in CASOS, and based on the knowledge about the impact of text coding choices on network data from chapter 2. Also, I have over six years of experience in using the text coding methods applied in this chapter. In future work, the validity of my findings should be further tested by other people by validating the auto-generated thesauri, master thesauri, and resulting network data.

3. Co-reference resolution: The main task for which these decisions and subject matter expertise were needed was co-reference resolution, which had to be performed in order to validate and refine the master thesauri and auto-generated thesauri, to refine the network data, and to clean the corpora. Since co-reference resolution on texts, thesauri and network data is not yet supported in AutoMap and ORA, I did perform these tasks by hand, which has limitations beyond the aforementioned time costs and risk of incompleteness, errors and biases: for example, I merged some nodes for which it was not perfectly clear if all instances of these nodes map to the same, specified real-world person (e.g "salva" to "kiir"). For these cases, I considered the entity frequencies (first name appears with similar or lower frequency than last name) and alternatives (merging only if no other agent with same first name or last name occurs in the union of the annual networks) to the best of my knowledge and limited subject matter expertise. For instance, in the Sudan data, some of the most frequent agent nodes were single names, e.g. "ibrahim" (5,822 instances) and "muhammad" (6,202 instamces). These could not be mapped with high certainty to more specific agents. In conclusion, the addition of co-reference resolution routines that operate on the network data level and the text data level (for thesaurus generation) would be a highly useful extension to this work. Such routines would need to be able to reason about the similarity of nodes not only based on string similarity, which would fail for cases like "Salva" and "Kiir", but also by exploiting external domain knowledge as well as structural features of the network data. Alternatively, conducting reference resolution on the input text data prior to generating thesauri would solve this issue in the same way as it is solved for meta-data networks, such that reference resolution is not pushed off to the thesaurus or network data level.

4. Link types: All approaches for extracting network data from texts used in this chapter treat links as untyped network constituents. Another valuable extension to this work would be the classification of links. In prior research, various ontologies for categorizing links between agents or organizations as "conflicting", "cooperative" or "neutral" have been developed and evaluated (Goldstein, 1992; McClelland, 1971). In political science, the categorization of links is a state of the art process in event data coding (Bond et al., 2003; Schrodt et al., 2008). Machine-learning based methods for predicting link types have also been developed (Bunescu & Mooney, 2007; D. Roth & Yih, 2002).

5. Link formation: The findings are limited by the link formation approach, namely windowing, used for the extraction of relational data from text data. The results in chapter 2 had shown that windowing involves the risk of false positive links. To further test the conclusions drawn from this chapter, the same tests should be repeated with alternative link formation methods.

6. Prediction models for thesaurus generation: The evaluation of the auto-generated thesauri per application scenario and entity class revealed multiple limitations that occurred in all scenarios. Based on the synthesis of these limitations as presented in the conclusion sections per application scenario in this chapter, I suggest exploring whether retraining the models with the following modifications would lead to more accurate thesauri in application scenarios:

- Train without the part of speech feature.
- Train with a lower iteration rate, e.g. 300.

- Add the classes that consistently perform low in the application scenarios to the "none" class.
- Provide more examples in the look up dictionary for the classes that consistently perform low in the application scenarios (Ciaramita & Altun, 2005; Cohen & Sarawagi, 2004).
- Use different, domain-specific look up dictionaries to train models for particular domains.

Yet another approach to achieve higher accuracy for the auto-generated thesauri without revising the thesauri for every new project would be to use more advanced domain adaptation techniques (Daumé, 2007; Gupta & Sarawagi, 2009; Satpal & Sarawagi, 2007). These techniques do not necessarily require the retraining of the prediction models, which can be a highly time-costly process, but use statistical techniques to adjust trained models to new domains.

7. Incompatibility between methods and tools: The insights are limited by a given technical constraint: the tools used herein for the D2M process and conducting network analysis convert all thesaurus entries to lower case and perform node comparisons on a lower case basis. On one hand, this work flow is consistent and coherent. It also is efficient, because it eliminates the need to add terms that typically occur in lower case, but occasionally appear capitalized, to thesauri in both forms. On the other hand, adjusting the thesauri so that they contain only lower case entries causes a considerable loss of information, such as the disability to differentiate capitonyms. An example are terms like Rice, Straw and Bush (people) and Turkey (organization) versus rice, straw, bush and turkey (generic natural resources); all of which would have been relevant for analyzing socio-cultural networks such as the Sudan data, but were typically reduced to the meaning that showed the higher frequency count based on the underlying text data. Another problematic example are the resulting incidental overlaps of key entities from networks constructed from the meta-data (wood as resource) and text bodies (wood as person): for these data, I hypothesize that differentiating between terms in upper and lower case form will show that author networks reconstructed from texts authored by these people are even smaller than those identified in this chapter. In future work, two strategies could be employed to mitigate this limitation: first, one could adjust the tools or use different tools for conducting analysis on a case-sensitive level. This strategy was beyond the scope of this thesis, but once implemented, the analyses conducted herein could be repeated in order to identify the qualitative and quantitative impacts of this change, and the robustness of the network data given an extraction method towards these changes. Second, part of speech, which are also output with high accuracy by the prediction models and in the auto-generated thesauri, could be used to disambiguate thesaurus entries and their matches in the text data. This would be particularly beneficial for distinguishing between proper nouns and common nouns (such as the examples shown above) and for eliminating a common type of error that the prediction models cause in the auto-generated thesauri: there, common nouns could be disregarded if the occur in upper case form, which happens at the beginning of sentence and is possibly due to the data sparseness at this position, often cause misclassifications as specific social agents and to a lesser degree also specific locations. This second strategy might be less effective than the first one, but is also less invasive in terms of changing existing technologies.

6 Methodology for Jointly Using Text Data and Network Data: Leveraging Social Groups to Advance the Enhancement of Social Networks with Content Nodes

Considering the content of text data pertaining to socio-technical networks is essential for understanding the effects of language use on networks, including the transformative role that language can play on networks, and the interplay and co-evolution of information on one side, and the properties and dynamics of networks on the other side (Bourdieu, 1991; Corman et al., 2002; Danowski, 1993; J. Milroy & Milroy, 1985; C. Roth & Cointet, 2010). One common way of fusing network data with information from text data is to enhance social networks with nodes representing salient information from the text data. An advanced version of this procedure is to extract both a social network and a semantic network from a corpus, and then merging these networks based on agents who are associated with the same pieces of knowledge and information (Harrer, Malzahn, Zeini, & Hoppe, 2007). Another example for such procedure are the multi-modal networks extracted in the previous chapter. However, this approach has been criticized for its arbitrariness of selecting content nodes to be linked to certain agents. In this chapter, I address this problem by leveraging research from network analysis on social roles, positions and groups to develop a methodology that allows for fusing social network data with relevant information from text data such that the selection of agent nodes to link to content nodes is informed by partitioning networks into structurally coherent group. This work helps to answer the following kind of question:

- How can prior research on social roles and groups be used to advance the enhancement of social networks with content nodes extracted from text data?

Based on a review of possible theoretical underpinnings for enhancing networks with content nodes, an interdisciplinary and computational solution to this problem is developed in the methods section of this chapter. This solution involves topic modeling, an unsupervised technique, for identifying salient text terms. In the operationalization and results section, this methodology is tested on real-world datasets that were previously used herein. Moreover, the proposed methodology is put into the wider context of this thesis by comparing the outcomes of topic modeling to the results obtained with the methods used in the prior chapter, including entity detection based on supervised learning. This comparison enables the addressing of the following kind of question:

- How does topic modeling compare to alternative information or relation extraction methods in terms of identified terms and themes?

6.1 Introduction and Problem Statement

When text data pertaining to networks are available as a source of information, people have several options for how to use the content of text data for network analysis. I have consolidated these choices into five methodological approaches, which are discussed below. This discussion concludes with the selection of one approach, for which I develop and test a resolution to the main limitation with this approach. In the context of this chapter, I distinguish between the content or substance of text data (text bodies), which have been written by people, versus meta-data, which can also contain text fields, e.g. index terms and key words, and can originate from human authors or algorithms.

6.1.1 Disregarding Text Data for Network Analysis

Even though text data are often acquired as a natural by-product of (network) data collection processes, this does not mean that these texts are necessarily useful or relevant for further analysis. Thus, if the content of text data does not contribute to the understanding of a network, the text data can be disregarded. Examples are the Funding and Enron datasets described in the previous chapter (5.3.1), for which explicit social network data (who collaborates or communicates with whom, respectively) were acquired together with the corresponding text data (abstracts of research proposal and email bodies). However, for conducting classic social network analysis, these text data might not be needed. Another argument in favor this strategy is a statement by White (1963, p. 5), who argued that the "distinctive aspect of roles in formal organization must be not their content but their articulation, the structure they form." Furthermore, disregarding text data for network analysis is the most efficient approach in terms of time costs discussed in this chapter. The main limitation with this approach is that to the best of our knowledge, there are no empirical studies that provide information on the conditions under which the consideration of text data for network analysis is useful or not, and how much of a difference in understanding a network it makes. Even though many methods and technologies are available for extracting network data form text data[22], what is missing here are decision support mechanism that help us to assess whether considering text data for a network analysis project will offer additional value or not. Even though this problem is not addressed in this chapter, the previous chapter has shed some light on this question; showing that:

- Networks constructed from meta-data hardly resemble ground truth data, while networks extracted from texts can partially lead to this effect.

[22] For a review of these methods see section 3.2, more elaborated reviews are offered in (Diesner & Carley, 2010c; Mihalcea & Radev, 2011).

The mutual resemblance of networks extracted from text data and meta-data is low in terms of nodes and minimal in edges, but networks extracted from text data still resemble meta-data networks better than vice versa. This could be a partial function of network size: networks extracted from text data using thesaurus-based entity detection and collocation-based link identification tend to be larger in terms of the number of nodes, edges, and node and edge classes than meta-data networks as well as network data constructed in collaboration with subject matter experts. Network size impacts the value of certain network metrics.

- For social networks, key entities from text-based networks allow for a more localized and domain specific view on networks than meta-data networks do. For knowledge networks, the inverse effect was observed: meta-data networks comprise more informative and descriptive key nodes, while the key nodes from text-based networks provide a more generic view on a domain and corpus.

6.1.2 Represent Content as Links

The content of textual information can be abstracted or reduced to the existence, weight or likelihood of nodes and edges. In the simplest and widely used version of this approach, any observed occurrence of the exchange of information between a pair of entities is be converted into a link, and the (weighted or scaled) frequency of these occurrence is used as the link weight (see for example Cataldo & Herbsleb, 2008; Diesner et al., 2005; Doerfel & Barnett, 1999; Gloor & Zhao, 2006; Haythornthwaite, 2001; C. Roth & Cointet, 2010). The main critique with this approach is that it may fail to consider relevant information about a network that is encoded in the text bodies (Alderson, 2008). Scholars in communication science, among others, have previously emphasized this limitation: Corman et al. (2002, p. 164) argue that we “cannot reduce communication to message transmission”. Danowski (1993, p. 198) states that “travelling through the network are fleets of social objects”, and capturing those objects requires the analysis of the text data.

A different instance of this approach, which is not subject to the abovementioned limitation, is the construction of directed influence diagrams about uncertain events. In these diagrams, subject matter experts denote events, the causal relationships between these events, and link weights that indicates the (estimated) likelihood of an event causing an effect (Howard, 1989; Pearl, 1988). This process is the basis for constructing probabilistic graphical models. A particular family of these models, namely conditional models, was used for representing dependencies between text tokens and node labels in section 3.3.

6.1.3 Analyze Text Data and Network Data Separately

The content of text data can be considered, but analyzed separately from the network data. This strategy is typically used to acquire additional information about nodes that have been identified as key entities with respect to certain network metrics. An example for this approach is link analysis, previously referred to as the production of Anacapa diagrams, where network data are generated as part of criminal investigations: once a network diagram has been constructed from evidence, hypotheses for further investigations are developed (Harper & Harris, 1975; Howlett, 1980). One method for testing these hypotheses is to go through the records and protocols collected on individuals. Another example is text analysis based on grounded theory methodology: there, human coders identify relevant concepts (codes), document the codes in memos, aggregate similar codes into variables, and arrange the variables into relational structures (Bernard & Ryan, 1998; Lewins & Silver, 2007). These relational structures represent the implicit relations in the data, and support the development of models and theories (Glaser & Strauss, 1967). All text passages that have been associated with a code or variable can then be retrieved, and in-depth, qualitative text analyses and close readings can be conducted on them. While this approach is suited for gaining a thorough understanding of sub-sets of documents and certain phenomena, the main limitation are that it does not scale up to larger text collections and often requires human interpretation (Corman et al., 2002).

6.1.4 Relation Extraction

When the structure and behavior of networks are encoded in the text data itself, network data can be extracted from the texts. This approach was discussed in detail in the prior chapters, but is mentioned here for completeness. Relation Extraction offers an alternative solution when reducing or abstracting the substance of text data to nodes and edges causes a loss of relevant information, and also when the entire text basis needs be considered for analysis in an efficient fashion. Once relational data have been extracted from a corpus, they can be used as stand-alone network data for further analysis or be combined and jointly analyzed with existing network data. For example, in the previous chapter, I had concluded that fusing meta-data networks with text-based networks allows for combining different views on a network.

6.1.5 Jointly Using Text Data and Network Data

There is a large body of literature from various disciplines that supports the argument that jointly utilizing text data and network data can lead to a more comprehensive understanding of networks (and texts) than exploiting either data source alone or in a disjoint fashion (Alderson, 2008; Bourdieu, 1991; Carley & Palmquist, 1991; McCallum, Wang, & Mohanty, 2007; J. Milroy & Milroy, 1985; Mohr, 1998; C. Roth, 2006). The problem here is that methods and respective tools for putting this goal into action are less well established (Dabbish et al., 2011; C. Roth & Cointet, 2010). I am focusing my discussion of this approach on the most widely used instance of it, namely enhancing network data with content nodes.

6.1.5.1 Network Enhancement with Content Nodes

The simplest yet powerful approach for integrating text data and network data is to enhance a network with nodes that represent pieces of content from the text data. I refer to these nodes as "content nodes", and to this approach as "network enhancement with content nodes". Content nodes typically represent salient terms from the text data. These terms can be found, for instance, by computing (weighted) term frequencies per (lemmatized) term, and picking the terms with the highest scores (C. Roth & Cointet, 2010). The content nodes are then linked to the agents who have generated, processed or disseminated the respective information. The resulting data can readily serve as input to regular network analysis methods (see for example Carley et al., 2007; Gloor & Zhao, 2006).

An example for network enhancement with content nodes is SmallBlue, an expert finder system that makes inferences based on the social network data about IBM's employees (Ehrlich et al., 2007). A study of SmallBlue has shown that enhancing social network data with information derived from people's blog entries, emails, chats, bookmarks, and other social media sources improves the systems' performance in terms finding experts (Ehrlich et al., 2007). This was particularly true when searching for experts on very specific, narrowly defined problems. In the previous chapter, an even simpler version of network enhancement with content nodes was used, where the social network of collaborators on research grants was connected to nodes representing index terms for these projects. These index terms are not from the actual text bodies, but are rather very general proxies for the content of the text data that were selected by the authors. In summary, network enhancement with content nodes is an efficient engineering solution that is easy to implement, and is widely and successfully used for practical purposes.

From a scientific point of view, the main critique of this approach centers on the arbitrariness of the process of identifying content nodes: first, the respective network enhancement process does not consider theories, models or prior knowledge about the relationship between the social position and role of individuals or groups in a network, or their language use (Corman et al., 2002; Woods, 1975). Consequently, connecting any one actor to content nodes happens independently from connecting other actors to content, even though it has been shown that social relationships impact the content that people produce, perceive and share, and vice versa (this relationship is discussed in more detail in the background section below). Second, the mutual influence of content networks or semantic networks and social networks is considered at most in one direction, i.e. the impact of social networks on concept networks, but not vice versa (Cowan, Jonard, & Zimmermann, 2003; Harrer et al., 2007; C. Roth & Cointet, 2010). This is problematic as there is prior research in support of the argument that without considering the content of text data, we are limited in our ability to understand the effects of language use in socio-technical networks, including the transformative role that language can play on networks, and the interplay and co-evolution of information and the structure and behavior of networks (Bourdieu, 1991; Danowski, 1993; Giuffre, 2001; J. Milroy & Milroy, 1985; Mohr, 1998).

To summarize the discussion of methods for considering text data and network data, I conclude that a) Relation Extraction and b) jointly using text data and network data are best suited for considering the substance of text data for network analysis if needed and useful. Relation Extraction has been addressed in the previous chapters. The focus of this chapter is on following research question:

> *Research question 1:*
> *How can the method of enhancing social networks with content nodes be advanced such that the arbitrariness of adding content nodes to social networks is mitigated?*

In the following background section, I discuss theories and prior work relevant for finding a resolution to the arbitrariness of adding content nodes to social networks. The main purpose with this chapter is to identify, implement and test a methodological advancement to the given method. Similar to the validation of the prediction models in the previous chapter, the resulting procedure is demonstrated in two application scenarios.

6.2 Background: Theories and Models for Jointly Using Text Data and Network Data

In this section, the concepts of social positions, social roles, and groups are reviewed with respect to the suitability of those concepts for informing the selection of agents to connect to content nodes.

6.2.1 Relationship between Social Positions, Social Roles and Groups in Networks and Language Use

6.2.1.1 What are social positions, social roles and groups?

In network analysis, the concept of *social positions* is defined as a collection of nodes that are similar in their activities, interactions and ties with respect to other positions (Breiger, Boorman, & Arabie, 1975; Burt, 1976; Wasserman & Faust, 1994). Thus, positions are equivalence classes. Conducting positional analysis basically means to identify, represent and analyze nodes partitioned into subsets. In each partition, the nodes are linked in similar ways to the nodes in other positions (Lorrain & White, 1971). This process is commonly referred to as grouping; with blockmodeling being a prominent example for a grouping method (H. C. White, Boorman, & Breiger, 1976). The outcome of positional analysis is a mapping of nodes to groups.

From a network analytic point of view, the concept of *social roles* is defined as patterns of relations between nodes or positions (Merton, 1968; Nadel, 1957; H. C. White, 1963). The focus with roles is on associations among relations that link social positions, not relationships between nodes. Furthermore, roles are not defined over pairs of positions, but on the network level, where roles describe how each pair of positions is related to each other. Individual nodes can have multiple roles. Furthermore, primitives of roles, e.g. the kinship relationships between descendants, can be combined into chains of roles or more complex roles, such as the descendant of a descendant, i.e. grandchild (H. C. White, 1963). The outcome of role analysis is a joint representation of identified positions (one node per position) and the relations between them. Common representations of this output are a) image matrices, where the nodes are positions and the values per cell denote the presence or absence of a connection, and b) reduced graphs, which are visualizations of image matrices (Wasserman & Faust, 1994).

Despite these formal, network-centric definitions of social positions and roles, theories about these concepts are often formulated in terms of the properties of (groups of) individuals (Merton, 1968). These properties can be structural ones (Lorrain & White, 1971; Winship, 1988) or other behavioral signatures: one example for *structurally defined roles* are the classic power roles developed in social network analysis, which are defined in terms on node level-centrality metrics as introduced in section 1.2.1 (Mandel, 1983). These power roles include brokers or gatekeepers (high in betweenness centrality), lobbyists (high in eigenvector centrality) and celebrities (high in degree centrality), among others. More recent examples for structurally defined roles are roles that express the exclusiveness with which nodes from certain node classes have access to nodes from other classes, such as the exclusive access of some agent(s) to resources and knowledge (Carley, 2002b). An instance of roles defined over *behavioral signatures* is homophily, which assumes that people who are similar in their personal characteristics tend to form links with each other, such that networks feature homogenous sets of people (McPherson, Smith-Lovin, & Cook, 2001). Further, research in anthropology has shown that the presence of people who play certain informal social roles in groups, e.g. the one of “expressive leaders” (people who organize social events, social directors) correlates with a cohesive group structures. At the same time, the absence of other informal roles, especially of “instrumental leaders” (people crucial for getting things done) is associated with fragmented groups (Johnson et al., 2003). Such empirically grounded insights about the relationship between roles and network structure are essential because the cohesion or fragmentation of a group is related to a) its performance (Krackhardt, 1994) and b) the potential for conflict in groups and their wider environment (Humphreys, 2005). Another example for a behavioral property that has been used to formulate hypotheses and theories about social roles is language use (Humphreys, 2005; Marcoccia, 2004; J. Milroy & Milroy, 1985). This point is elaborated in detail in the next section (6.2.1.3).

Two closely related areas where fundamental theories about network positions and roles were developed are the adoption and diffusion of innovation as well as opinion leadership (Coleman et al., 1966; Rogers, 1962; B. Ryan & Gross, 1943). In these areas, roles mainly comprise innovators, early adopters, different types of majority, and laggards, and also the concept of boundary spanners. These role models have been adopted and further advanced across disciplines (Burt, 1999; Katz & Lazarsfeld, 1955; McAllister & Studlar, 1991; K. H. Roberts & O'Reilly III, 1979; Tushman, 1977), and also been tested for their current applicability to social behavior (Watts, 2007). Currently, role analysis is also a heavily researched topic in social media analysis. For example, roles that individuals occupy in discussion forums and learning systems have been identified by analyzing the structural position of individuals in a graph (Stuetzer, Carley, Koehler, & Thiem, 2011; Welser, Gleave, Fisher, & Smith, 2007) as well as the text data provided by network participants (Golder, 2003; Haythornthwaite & Gruzd, 2008).

In general, the underlying assumption with all network-oriented research on social positions and roles is that the identified patterns in observed relations are indicative of the roles that nodes in different positions play. There are many theories about the relationship between node properties and positions and roles, which is mainly due to the following reason: “since there are numerous ways to formalize the idea of types of ties, there are numerous ways to formalize the ideas of network role and network position” (Wasserman & Faust, 1994, p. 464).

In summary, due to the less strict definition of roles in theories about networks and human behavior, roles are not only specified and therefore operationalizable on the (global) network level, where the definition of roles is typically rather abstract (Wasserman & Faust, 1994; H. C. White et al., 1976), but also on the local level, i.e. on the level of nodes and positions (Mandel, 1983; J. Milroy & Milroy, 1985; Sailer, 1979; Winship, 1988). This review has furthermore shown that theories about social positions and roles often originate from the consideration of structural and other behavioral characteristics of (groups of) nodes; with one of these features being language use.

6.2.1.2 General concept of groups

Social positions and roles are a particular instance of groups that can be identified from graph data. Zooming out from the specific level of positions and roles to a more general level, groups represent sets of nodes that are structurally similar to each other (Wasserman & Faust, 1994). A commonly used alternative to the notion of structural equivalence, i.e. roles and positions, is the idea of groups defined by cohesion. Simple forms of cohesive groups that have been previously introduced in this thesis are triads, cliques and components (Table 154, (Krackhardt, 1998; Wasserman & Faust, 1994)). More elaborated notions of cohesion involve partitioning a graph based on network properties of nodes and links, such as the betweenness centrality of nodes (Girvan & Newman, 2002). The main difference between groups defined by structural equivalence versus cohesion is that in the first category, group members might be dispersed over disjoint or distant parts of the network, which is not the case for group members from the second category.

6.2.1.3 How do social positions, roles, and groups relate to language use?

What do we gain from considering texts *and* networks over using only either one data source? Research on language change has shown how the network position or group membership of social agents is indicative of the social roles that people or groups play with respect to language change (Gumperz, 1982; Lippi-Green, 1989; J. Milroy & Milroy, 1985; L. Milroy, 1987). The Milroys have found that boundary spanners who adopt new facets of their vernacular are most effective in spreading these changes into the wider community. More specifically, the structural properties of people who are effective in introducing and diffusing innovation are a plethora of weak ties (for the notion of strong and weak ties see Granovetter, 1973), marginality to any adopting group, and an attitude of not considering the elements of change as a significant network marker. In contrast to that, people who are located at the core of cliques and hubs can afford and in fact tend to resists to impacts that deviate from the group's norms and that originate from outside their group. This area of research has concluded that people's attitude towards language change impacts greater sociolinguistic patterns of the adoption and diffusion of vernacular. For some of this work (J. Milroy & Milroy, 1985; L. Milroy, 1987), multiple types of ties have been considered, namely kinship, friendship, collaboration, and being neighbors, which illustrates the point that the analysis of roles and positions is more informative if multiplex data are used (Wasserman & Faust, 1994; H. C. White et al., 1976).

Work by Eckert (2000) has shown how in groups that are formed for a certain purpose (communities of practice), linguistic styles are continuously developed and shared by the group members. Consequently, the homogeneity of language use in such groups increases over time. This work ties back to the concept of homophily (McPherson et al., 2001). Also investigating the convergence of language, Fitzmaurice (2000) used historic data (letters) to investigate how strategies alliances between individuals impact their language use. She showed that in the contexts of hostile or competitive situations, people who may have opposing agendas but a shared goal, form dense network clusters. In these groups, language use becomes more homogenous. There is also support for the reversal of this effect: We have shown how during an organizational crisis, the entropy of the content of interpersonal communication decreases, while polarization increases (Diesner, Carley, & Katzmair, 2007).

Guiffre (2001) discovered a positive relationship between the stylistic perceptions of artists as expressed in reviews written by art critics and the decisions made by gallery owners about concurrently exhibiting work by different artists. The more favorable the reviews for any two artist were, they more likely it was that they were co-exhibited. This relationship is self-reinforcing over time; ultimately leading to more or less successful careers in art.

Roth and Coinet (2010) found that the relationships between social capital, measured as degree centrality of authors, and semantic capital, operationalized as highly central documents, differs depending on the type of collaboration that a group in involved in: for scientists who co-publish, social capital and semantic capital show a significant, positive covariance. For contributors to social media (bloggers), a different trend was observed: poor semantic capital does not translate into low social capital, i.e. authoring non-popular or marginal comments does not hurt a person's social status.

In summary, prior work from different areas has provided empirical evidence as well as a few theories and models for the relationship between language use and the membership of people in groups in networks. Also, this review has shown that jointly utilizing texts and networks requires interdisciplinary work at the intersection of natural language processing, network analysis and maybe other fields, especially sociology and anthropology. While this intersection still forms a small yet growing area of research, no commonly accepted methodology for putting this idea into action has yet been adopted. In the next section, I build upon prior work in natural language processing and artificial intelligence to develop such a methodology that integrates prior knowledge about groups with an efficient, non-arbitrary method for identifying content nodes that also are grouped into sets of similar entities.

6.2.2 Roles, Positions and Groups at the Text Data Level

The idea of positions, roles and groups has also been conceptualized for the text level. I focus my review of prior work in this area on research related to network analysis. Partitioning words into groups of similar or equivalent sets has a long tradition in network analysis: initially, researchers have mainly used multi-dimensional scaling (MDS) as a method to this end (Woelfel, Holmes, Cody, & Fink, 1988). MDS basically transforms a squared matrix into Euclidean distances between nodes (Kruskal, 1977). The output of this process is a two-dimensional, graphical representation of the proximity between any pair of nodes. The assumption with or interpretation of this semantic space is that the closer two nodes are, the stronger is their contextual semantic association. Especially in communication science, MDS has been used to cluster words from documents (Doerfel & Barnett, 1999; Woelfel et al., 1988), and also to partition communication networks into groups of participants who are similar in their communication behavior (W. D. Richards, 1971; W. D. Richards & Rice, 1981). Another methods that can be used for partitioning words is Latent Semantic Analysis (LSA); also referred to as Latent Semantic Indexing (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSA is based on the same matrix operations and underlying assumptions as MDS, and has also been used for practical applications of grouping words (Smith & Humphreys, 2006). In LSA, Principal Component Analysis (PCA) is applied to word-document co-occurrence matrices, and the output is also a two-dimensional representation of word or node proximities.

There are three main disadvantages with the spatial models described above (Griffiths et al., 2007): first, the identified relations are always symmetric, even if they are truly asymmetric. For example, a stalker feels closer to his victim than vice versa. Second, these models do not allow for term disambiguation, because all semantic associations of heteronyms appear in equal proximity to the focal concept. Consequently, unrelated terms can be placed into the same position. Third, these models can wrongfully suggest coherent local substructures (groups) such as triads or cliques. For example, politicians might be friends with trade union leaders and business executives, which does not imply that the trade union leaders are also friends with the business executives.

An alternative model that also takes document-word co-occurrence matrices as input and outputs terms grouped into positions is topic modeling; a technique based on Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). In contrasts to MDS and LSA, LDA is based on the assumption that there is a probabilistic, generative process according to which some assumed latent, unobservable structure generates words, which can be observed. Bayesian inference is then performed on the observed words to infer the latent structure. The specifics of the assumed latent structure and the causal (generative) dependencies between the considered variables can be expressed as probabilistic, graphical models. Typically, topic models are represented with plate notation.

The commonality between MDS, LSA and LDA is that these techniques are unsupervised techniques that basically reduce the dimensionality of text data to unlabeled sets of terms that are related through their context-specific, semantic associations (Griffiths et al., 2007). In topic modeling, these sets are called topics. In contrast to MDS and LSA, LDA can disambiguate between different meanings of a word (the same term can appear in multiple topics). Moreover, LDA does not enforce symmetric relationships or triads and closures of larger node groups. In topic modeling, each topic comprised a set of words where the weight per word indicates the strength or likelihood of the association of a word with a topic. The assignment of words to topics is a non-exhaustive and non-exclusive process, meaning that not all text terms are descriptive for topics, while certain terms or phrases may occur in multiple topics. Topic modeling has become a state of the art technique for grouping words into sets that express the gist of a corpus. To a lesser degree, topic modeling has also been used in the context of network analysis (Diesner & Carley, 2010b; McCallum, Wang, & Mohanty, 2007).

Another approach to grouping words is based on the theory or assumption of spreading activation. This approach assumes that mentioning a concept triggers the activation of semantically related concepts, which can be retrieved from human or electronic memory (Collins & Loftus, 1975; Collins & Quillian, 1969). Translating this idea into network analysis terminology means that a concept is defined by its ego-network. An ego-network comprises all nodes in the one-step environment of a node, such that the size of the ego-network equals the node degree (Carley, 1997a, 1997b; Mohr, 1998). Since spreading activation uses a similar data structure or representation for nodes and edges like MDS and LDA do, this approach also suffers from the inability to disambiguate identically spelled terms with different meanings.

Finally, Carley and Kaufer (1993) have proposed a typology for grouping concept nodes from semantic networks into eight ideal types or primitives that describe the communicative connectivity and communicative power of nodes. Nodes are assigned to these types based on their combined score on three dimensions: density (total node degree), conductivity (betweenness centrality) and consensus (frequency of ego-network of a node). For example, words scoring high on conductivity, but low on consensus and density are categorized as "buzzwords". Only extreme values on these dimensions ("high", "low") are considered, such that the grouping process is not necessarily exhaustive. This approach has a limitation that generalizes to automated methods for grouping words based on their value for network metrics in general (Diesner & Carley, 2010b): the magnitude or range of these values have no absolute, predefined or theoretically grounded interpretations, such as a density of 0.2 could be interpreted as being high, low or medium. Instead, most of these metrics can only be interpreted in comparison to the values computed on other networks or the same network at another point in time. Therefore, appropriate cut-off values for determining when a node scores high or low on a metric can only be defined as rule sets or heuristics. This requires a data-driven, case-wise decision-making process, and also a basic understanding of the network metrics. The resulting limitation is that this approach to grouping nodes cannot be fully automated, and moreover does not generalize from one dataset to another one without testing the appropriateness of cut-off values and potential adjustments (Diesner & Carley, 2010b). Consequently, this process is expensive in terms of time and human resources.

6.2.3 Summary of Insights Gained from Review of Theories, Models and Methods for Jointly Utilizing Text Data and Network Data

Summarizing the insights from this review section leads to the following conclusions:

1. The approach of enhancing network data with content nodes is practical and efficient. However, the identification of content nodes is arbitrary and lacks a theoretical foundation. Also, the mutual influence of network data and language use cannot be appropriately considered.
2. The limitations named in the paragraph above can be alleviated by drawing from the rich body of previously developed theories, models and methods for grouping nodes (social actors, other socio-technical entities, and words) into structurally similar network partitions. Two notions of groups were discussed:
   - Groups defined in terms of equivalence classes (social positions) and relations between these positions (social roles). In contrast to the initial strict definition of roles and positions and due to theoretical and methodological advances, analysis of roles and positions can be conducted not only on the network level, but also on the level of nodes and node groups.
   - Groups defined by cohesion.
3. Topic modeling is an efficient and appropriate technique for grouping words.
4. Jointly considering groups of nodes (social or socio-technical network) and text data for network analysis has led to insights that cannot be gained by using either data source alone.

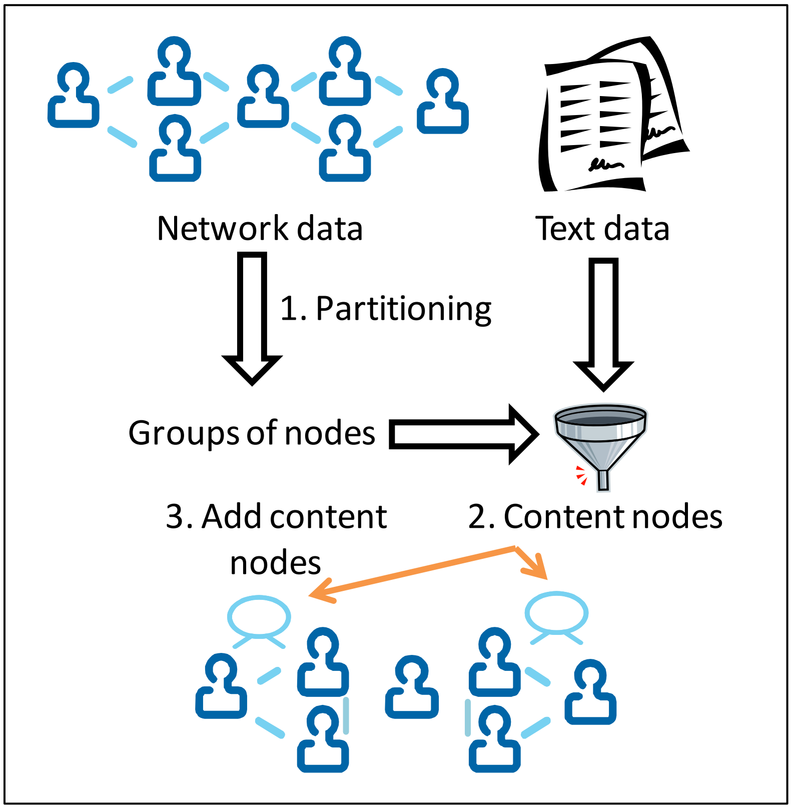
6.3 Methodology

The first research question in this chapter was: How can the method of enhancing social networks with content nodes be advanced such that the arbitrariness of adding content nodes to social networks is mitigated? In this methods section, I develop an answer to this question by turning the conclusions made above into a three step methodology that where the selection of a) agents to be linked to content nodes and b) the identification of content nodes are non-arbitrary. Figure 14**Error! Reference source not found.** illustrates the proposed workflow. Steps one and two require decisions or strategies for operationalizing the grouping of actor nodes and the selection of content nodes. Step three is a straightforward and a deterministic matrix operation.

Therefore, I focus the following section and subsequent analyses on steps one and two, and provide a user guide for conducting step three.

1. Partition social networks into groups.
2. Identify content nodes per group. This step serves the identification of shared content per group. One option is topic modeling on the texts originating from the nodes per group.
3. Enhance social network with content nodes.

Figure 14: Workflow for proposed methodology



6.3.1 Partition Networks into Groups

The first question for this step is: What social positions, roles or groups to consider? Wassermann and Faust (1994) recommend to use rather general and abstract conceptualizations of the structural location of nodes in networks when formalizing social positions and roles, and also to use flexible descriptions of patterns or types of relations between nodes. Our prior research supports this recommendation: we had identified and compared the content produced by people who occupy roles that represent their disposition and ability to motivate or inhibit language change in social networks (Diesner & Carley, 2010b). These roles were based on empirical work and a resulting theory by Milroy and Milroy (1985; 1987). Being in the position to change or maintain norms in a group and possibly also in the wider society bears opportunities and risks for members of either group. In order to assign nodes to these two groups, we had developed role templates that combined multiple node-level network metrics that we evaluated as being are relevant for detecting particular roles. Then, we identified nodes that fit either template by computing the selected metrics on all members, and screening the results to define boundary or cut-off values for scoring high, medium and low on each metric. Finally, we performed topic modeling on the set of all texts per group. In the context of this chapter, there are limitations with this approach: First, it cannot be fully automated, and therefore does not scale up. This is because there are no predefined, logical or empirically or theoretically grounded values that are indicative of scoring low, medium or high on certain network metrics. Therefore, these boundaries have to be manually identified on a per group basis. Second, this approach does not generalize across networks, which is for the same reason as the first issue. This means that for each network or time slice of a network, group membership has to be identified separately. Third, our prior approach was designed for a different purpose, namely comparing the language use of certain roles in order to answer the following research questions: What topics are addressed by members of each group? Which topics are exclusive to a group, and which ones are shared among groups? We argued that for this purpose, the method is useful. However, in this chapter, the focus is not *comparing* the language use or content of groups, but on facilitating the identification of concept nodes for the *enhancement* of network data. For this process, the following goals were identified in the background review section of this chapter:

First, identifying concept nodes not in an arbitrary fashion, but based on structural properties of the nodes that have generated, disseminated or processed the respective content. These nodes are typically social agents, i.e. individuals and organizations, and possibly also automated agents. For simplicity, I herein refer to them as agent nodes.

Second, adding concept nodes (here referred to as the knowledge network) to agent nodes (here referred to as social network) such that the agents are linked through content nodes, regardless of whether these agents already share a link or not. In this context, using our prior approach of identifying structurally equivalent agents implies the following limitations: taking the Funding data as an example, nodes representing the roles of formal leaders, for instance, might originate not only from different areas of the network, but also from different research domains (e.g. physics, economics). Comparing their text data within and across roles helps us to identify in what areas or on what topics these people are working, how they focus their proposals on terms related to project management terms or the subject matter domain, etc. – all of which are instances of role comparisons. However, it does not seem reasonable to link these agent nodes to shared content nodes since it is unlikely that leaders from different fields share any content beyond generic project management terms and terms indicating the potential for leadership, excellence and innovativeness. In fact, our prior research has shown that the strongest topic for the considered roles was project management; confirming the limitation outlined above. The same effect can even occur within a research domain, i.e. leaders emerge around different sub-fields. Another risk with linking people within a structural equivalence class is that agents could get connected to content nodes or knowledge that they were never truly exposed to, but that were simply salient in disjoint or distant parts of the overall network. In summary, enforcing knowledge nodes onto agents this way entails the risk of false positives. In conclusion, for the purpose of enhancing social networks with content nodes, it seems more reasonable to only link agent that could get exposed to the same content. Therefore, the next question is: Which grouping algorithm to employ? This question is answered in the results section of this chapter based on empirical tests in application scenarios.

6.3.2 Identify Content Nodes per Group via Topic Modeling

Topic modeling has the following properties, which help to overcome several of the aforementioned limitations of alternative approaches for extracting themes and salient terms from knowledge networks (Griffiths et al., 2007). Note that the input to topic modeling are document-term co-occurrence matrices, which can be considered a type of knowledge network:

1. Efficient: since the learning is unsupervised, no labeled ground truth data is necessary to build a prediction model. Also, no thesauri need to be constructed.
2. Scalability: Scales up to large corpora.
3. Word sense disambiguation: can identify different meanings of a word by considering the word's context.
4. Assumed generative process: the way topic modeling is operationalized here is based on the following assumptions: groups of people generate documents by selecting topics from a pool of topics, and words per topic from a pool of words. This generative process is probabilistic, but not arbitrary.

With respect to property one, there is a lack of knowledge about the following question, which aligns with the kind of research conducted in the previous chapter:

> *Research question 2:*
> *How do key entities identified by applying certain prediction models trained with supervised learning compare to entities identified via topic modeling?*

I provide an answer this question in the results section.

Topic modeling has been linked to network analysis before: Chang et al. (2009) have used the LDA) technique to suggest link labels for untyped links in semantic networks. McCallum et al. (2007) have conducted topic modeling on all bodies from two email datasets, and compared the resulting groups of people who are involved in the same topics. They conclude that identifying equivalence classes of people via topic modeling returns more reasonable groupings than using classic grouping methods from network analysis, and also better groupings than an alternative method for applying topic modeling on co-authored documents (Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004).

Mimno and McCallum (2008) argue that while in the basic version of LDA, any observed and descriptive features of the text data are generated based on an assumed latent probabilistic graphical model, conditioning topics on the observed data instead of generating the data might be more efficient. Based on this rationale, they developed the Dirichlet-Multinominal Regression (DMR) technique as an extension to LDA. The key idea with DMR is the assumption and computation of distributions per topic not only over words, but also over meta-data that provide additional information about documents. Thus, DMR eases the consideration of various types of meta-data on the text data, such as the date or publication venue of a text document.

In this chapter, I am drawing from the work mentioned above. However, with the proposed methodology, I am not learning a topical profile per individual, dyad or document, as done in prior work, but create topical profiles conditioned on groups. Moreover, I show how the themes and terms identified with topic modeling compare to the outcome of alternative methods for extracting this information, including supervised learning. As points of comparison, I am re-using the methods that were introduced and applied in the previous chapter, including supervised learning for entity extraction. The advantages, limitations and some typical results with respect to these methods (applied to the same data in the previous and current chapter were already identified in the previous chapter. Moreover, comparing these methods to topic modeling helps to put the outcome of this chapter into the wider context of understanding how different information and relation extraction methods relate to each other, and what different views on a network they can provide.

6.3.3 Enhancing social network with the content nodes

The top N content nodes are linked to the members of the respective group. In the case of social networks, the content nodes are added such that a two-mode, agent-to-knowledge network is created. Section I in the Appendix provides a step-by-step guide for operationalizing this procedure in ORA.

6.3.4 Evaluation of Content Nodes identified with Topic Modeling

One main limitation with topic modeling is the evaluation of the outcome: while the underlying, probabilistic graphical model as well as the overall method for performing topic modeling are clearly defined, the interpretation of the resulting topics is not a standardized process. This interpretation leaves plenty of room for making sense of the outputs or reading meaning into them (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009). In prior work on advancing topic modeling, such as adding new parameters on which the generating of words in constrained on, people have often used datasets that they were intimately familiar with, such as their personal emails, or data that are easy to interpret, such as news wire corpora. While this is a legitimate strategy, the following questions often remain unanswered:

> *Research question 3:*
> *How does topic modeling compare to alternative information or relation extraction methods in terms of identified terms and themes?*

6.4 I develop an answer to this question by comparing the outcome of topic modeling per group to the content nodes obtained by using the relation extraction methods that were evaluated in the previous chapter. While this step is not part of the proposed methodology for enhancing social network data with content nodes, it contributes to the validation of the outcome of topic modeling in a comparative fashion. Operationalization and Results

The proposed methodology is designed for enhancing datasets for which both social network data as well as text data are available. This applies to the Funding corpus (for details on this dataset see 5.3) and the Enron corpus (5.4). I also discuss the applicability of the methodology to the Sudan corpus, which contains text bodies and non-relational meta-data.

6.4.1 Application Context I: Sudan Corpus

The presented methodology is designed for situations where both social network data and text data are available, and is therefore not applicable when only either one data type is available. In contrast to the Funding and the Enron corpus, the Sudan corpus contains only text-data and non-relational meta-data, but no social network data. However, it has been shown in the previous chapter how social network data can be extracted from text data (5.2.2.1, 5.2.2.2) or constructed from meta-data (5.2.2.3). Using these network data as a proxy for explicit social network data, I tested the proposed methodology on the agent networks extracted from text bodies (as described in 5.2.2.2) and constructed from meta-data (as described in 5.2.2.3). Then, I applied the Girvan-Newman grouping algorithms to these networks. The resulting main groups contained agent nodes consistent with the key players identified in 5.2.3, i.e. political leaders from the Sudan, neighboring countries, and the Western world. Since the Sudan corpus contains no texts authored by these people, as a proxy, I retrieved all texts in which these people occurred. This resulted in large text sets per group, which had the following limitations:

- High overlap in text documents between groups. This is likely to lead to similar topic modeling results across groups.
- Texts per group co-mentioned many additional agents that were not part of the set of identified key entities.

Besides these issues, there are further limitations with modifying the proposed methodology to make it work for text data only. First, the modified methodology is circular in that it extracts social networks from the same pool of texts that are filtered for text bodies to process with topic modeling later. Second, the selection of content nodes to associate agents with is still fairly arbitrary because it is not necessarily the case that any agent mentioned in a document is associated with any content node also found in this document. A more reasonable approach to linking nodes representing agents to nodes representing pieces of knowledge or information that these agents are more likely to be associated with is extracting multi-mode network data as done in the previous chapter (5.2.2.1, 5.2.2.2). Third, if social networks data distilled from text data are used, all limitations with relation extraction step (partially addressed in chapters 2 and 5) will propagate to the grouping and text selection steps. Consequently, the findings could be impacted by this process, and empirical tests would be needed to identify the magnitude of these impacts. For these reasons and based on the described pre-tests, I decided to not further test the proposed methodology on the Sudan corpus.

The conclusion from this application scenario is that the proposed methodology is not suitable for datasets that do not entail explicit social network data. This might apply to collections of news wire data and other collections that consist only of unstructured, natural language text data. Furthermore, this argument is independent of the grouping algorithm. However, the grouping technique might impact the key players one would find. As an alternative method for combing social network data and semantic network data as contained in the text bodies themselves, one could use the multi-modal relation extraction techniques that were explained and tested – also on the Sudan data - in the previous chapter.

6.4.2 Application Context II: Funding Corpus

6.4.2.1 Social Network Data

For the social network, I used the collaboration networks that I created from the explicit denotation of which people were jointly funded for a grant. The construction of these networks is described in detail in section 5.3.2.3. Given the various levels of completeness of the social networks per framework programme (FP) (Table 105) and the respective limitations as explained in 5.3.2.3, I use the networks of FP 4 to 6 for this study. The collaboration networks are weighted, directed graphs.

6.4.2.2 Grouping of Social Network Data

In order to identify useful (with respect to the methodology) groups in the social network data, I tested various grouping algorithms as implemented in the ORA. Several of these algorithms did not return results on these sizable networks (Table 105) with a decent number of groups (about 10) in a reasonable amount of time. Since the goal here is not to find an exhaustive grouping of all nodes, I reduced the social networks as follows: first, I dropped all pendants, which are nodes that are linked to only one other node. Also, multiple pendants can be connected to one and the same node; resulting in marginalized power structures that may exhibit norm enforcing behavior. Pendants can be considered as a structurally equivalence class of their own that represents a certain role, i.e. the one of dependents. Next, I removed the resulting isolates. One caveat with this approach is that the last two steps eliminate any project team of size two. At this point, the network data were still too large for grouping. Therefore, I reviewed the node degree distributions, which followed the skewed distribution that is typical for social networks, and based on this review also dropped nodes with a frequency of one. Finally, I removed the resulting isolates again.

CONCOR is a classic grouping method that basically correlates the adjacencies between nodes in an iterative fashion (Wasserman & Faust, 1994). This technique is a parametric method which requires the specification of the number of groups to find a priori. Visualizing the resulting groups revealed that with CONCOR, the largest group mainly contains the collaborators on two-person projects (those that remain or result from the abovementioned graph modifications). The second largest group mainly comprises PIs on two-person projects. The third largest group are collaborators on three-person projects, and the fourth largest one are the PIs on three-person projects. This pattern continues. These groups clearly represent meaningful structural equivalence classes. However, as discussed above, it does not seem useful to perform topic modeling on the texts per group to identify shared knowledge, since these texts might have little in common beyond the dependency structure of their contributors. Thus, the knowledge represented in these texts is highly unlikely to be shared among the group members.

The same argument applies to groups identified based on key entity analysis: I computed the same metrics as in the previous application scenario for the Funding data 5.3.3) on the social network, and identified the top ten agents with respect to these metrics. Visualizing the resulting network suggests that the key entities are dispersed across the graph with little cross-connectivity among them. This point further supports the previously raised concern that structurally equivalent nodes might be exposed to disjoint pieces of information.

As another alternative, I used the Girvan-Newman grouping algorithm (Girvan & Newman, 2002). This algorithm basically identifies groups with strong internal connectivity, but weak connectivity to other groups. This is achieved by iteratively dropping edges with high betweenness centrality. Girvan-Newman is a non-parametric method, i.e. the number of groups to find must not (but can be) pre-specified. The fundamental difference between this algorithm and the previous two grouping strategies is that Girvan-Newman mainly forms groups of nodes that can reach each other within a few steps. Based on the discussion in the methods section, this property is desirable for this project because nodes that are separated by a few links are more likely to get exposed to the same content than nodes that might have perfect structural equivalence, but are located in disjoint components of the networks. Visualizing the resulting groups suggested that the identified groups seem appropriate for this study and dataset.

As a logical follow-up on the Girvan-Newman grouping algorithm, I also tested grouping based on components, which are disjoint section of a network (Table 154). The same advantage as pointed out for Girvan-Newman also applies to component-based grouping: nodes within a component have a higher chance of getting exposed to the same information by either working on a grant together and/ or via information diffusion through the wider network than structurally equivalent nodes from different components. Visualizing the component-based groups showed that they are very similar to the ones found with Girvan-Newman, and even are often identical for small groups (about ten members and less). The difference is that Girvan-Newman occasionally finds sub-groups within large components, which are less deterministic than the groups just based on components.

In summary, considering the limitations and advantages outlined in this section together with the requirements and goals for the proposed methodology, we decided to use the Girvan-Newman algorithm for grouping social networks.

Table 120 shows the number and size of groups obtained per FP considered. Across all FPs, most groups have a size of two. Many of these groups are actual project teams, where the members are involved in the same proposal. For this study, I am focusing on less deterministic groups that may and in fact in many cases do involve multiple funded research projects.

Table 120: Number and size of networks and groups

Dat a	Raw			Groups			Number of groups					
	Nodes	Edges	Texts	Nodes	Edges	Modula rity	Num ber	Min	Max	Aver age	Std Dev	10+ nodes
FP4	35,061	34,583	9,651	373	262	0.97	120	2	21	3.1	2.8	5
FP5	34,541	48,670	12,669	1016	1118	0.80	188	2	147	5.4	13.4	13
FP6	39,848	43,033	9,184	649	441	0.99	210	2	13	3.1	1.9	3

6.4.2.3 Identify Content Nodes per Group via Topic Modeling

For each FP and each identified group, I extracted all proposals that each member of the group was a PI on. This can entail proposals that group members have co-authored with others outside the group. I made this design choice to account for the possibility that the group might still benefit from this knowledge or this knowledge might diffuse through the group.

LDA takes term by document matrices as an input. In order to generate these matrices, I performed semantic network extraction in AutoMap by considered all tokens as concepts except for the entries specified in the delete list used in the previous chapter as well. For link formation, I used the windowing technique with a window size of seven. The windowing technique and choice of window size were also explained in the previous chapter.

Next, I conducted topic modeling on the semantic networks in ORA: I ran pretest with different numbers of topics (5, 10, 20), and based on that decided to use ten topics for FPs 4 and 6, and 20 for FP5. The differences in numbers are largely due to the differences in the amount of documents per groups per FP. Additional parameters that need to be set in ORA relate to the Gibbs sampling method. In consultation with Aparna Gullapalli, who developed the LDA implementation for ORA, I initially selected the following parameter values: step size: 100, iteration rate: 2,000, beta-value: 0.5. Inspecting the resulting topics showed that many of them involved numerical values, which seemed mainly noisy. Therefore, I re-generated the semantic networks as described above, but removed all numericals from the data. Inspecting the networks again revealed that multiple runs with the same parameter configuration returned different topics and topic members. This is no surprise since Gibbs sampling is a probabilistic method that uses random seeds, so that results may vary across runs. However, with a sufficiently larger iteration rate, the membership probability per topic should converge to some degree. I further explored this issue by increasing the number of topics to 30 and the iteration rate to 5,000. I used this modified configuration (the other parameters were kept constant and at the values named above) to perform three topic modeling runs each on a small, a medium size and a large semantic network from the Funding data, and compared the results across runs per network. This process confirmed the previous observation, i.e. that topics and members differ more strongly than one would reasonably expect across runs with identical parameter settings. Table 121 shows an example for the first five topics for a small network with an iteration rate of 2,000. There, the green cells indicative duplicate entries from different runs – what I was hoping for here is a high amount of green cells per run. While robustness of topic modeling is no requirement for the proposed methodology, some coherence in results produced across runs is needed for two reasons: first, to overcome the arbitrariness of finding content nodes, which is a limitation of alternative methods for enhancing social networks with content nodes. Second, to ensure the reproducibility of the results presented herein. One solution to this problem would be to change the implementation in ORA to use variational EM instead of Gibbs sampling. Another solution would be to work with a different existing implementation. I decided to test whether LDA-based topic modeling as implemented in the Mallet package leads to more robust results (McCallum, 2002). Table 122 shows the top five topics for the same network as used for Table 121. To produce these results, I generated ten topics with ten members. The results indicate two things: first, there is a higher overlap of topic membership (green cells) across runs on the same data with the same parameters with Mallet than with ORA. Second, LDA in ORA and Mallet retrieve very different themes and terms. The results from Mallet suggest that the text data are about transportation and policy, while with ORA, it seems hard to identify an overarching theme (topic label) for the retrieved terms. However, without any solid validation based on ground truth data, it cannot be said which implementation retrieves more appropriate results. All that can be concluded from this small-scale comparison is that the results from Mallet are more robust. For this reason only, Mallet was used for further analysis. Finally, I tested various numbers of topics to generate with Mallet (10, 20, 30, 50), and decided to stick to the initial number of ten for FP4 and FP6, and 20 for FP5.

Table 121: Topic groups for FP4, node group 1 (LDA in ORA)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Run 1	urban	chains	concentrates	investments	compared
	investigated	derive	ddg	barrier	dysaf
	co-ordinated	inter-operability	east-west	covering	consideration
	innovations	foresee	calibration	addressing	developed
	co-ordination	innovations	impulse	behaviours	appended
	20040101…	maintenance	bottlenecks	rail-ten	measure
	auspices	purpose	draw	interfaces	ballasted
	corridors	assist	sensitivity	20040101_14…	backcasting
	links	defining	urban	degree	apricot

	handbook	allowing	contribution	contradictory	
Run 2	professional	eastern	20040101…	compete	criteria
	conduct	databases	forms	central	players
	ground-based	fulfilment	documented	disseminated	varying
	derive	links	aim	seagoing	margin
	nox	corresponding	collected	allow	bundles
	operated	meet	arrangements	20040101…	20040101…
	easy-to-use	effect	aims	foresee	axes
	deliverable	preliminary	observatory	temporality	fifth
	committee	degradation	maintenance	harmonisation	advanced
	found	rd	conceive	centres	structures
Run 3	low	extended	covering	bft	appendices
	issue	effect	alps	aggregation	databases
	efficient	fasteners	prototype	aimed	freight
	sensitivity	consistent	20040101…	20040101…	commission
	evident	applicable	20040101…	acceptance	describe
	calibrate	devoted	by-road	analyze	southampton
	competitive	produces	collecting	corridors	follow-ups
	examples	eastern	track	bridges	integrates
	lisbon	capacities	unfold	co-operation	core
	accessibility	deliverables	administrations	deals	intermodal

Table 122: Topic groups for FP4, node group 1 (LDA in Mallet)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Run 1	policy	transport	transport	transport	intermodal
	methodology	project	scenarios	road	transport
	projects	european	development	freight	quality
	strategic	system	study	sea	freight
	assessment	economic	pricing	costs	market
	research	cost	relevant	project	actors
	european	freight	work	infrastructure	decision
	develop	infrastructure	eu	systems	services
	transport	improvements	countries	european	traffic
Run 2	policy	transport	transport	transport	intermodal
	projects	european	data	freight	transport
	programme	system	scenarios	sea	project
	assessment	project	methodology	project	chains
	strategic	economic	mobility	costs	quality
	research	interoperability	evaluation	urban	freight
	transport	research	pilot	services	traffic
	methodology	infrastructure	model	european	examine
	european	freight	demonstration	infrastructure	case
Run 3	transport	programme	policy	data	intermodal
	project	research	task	transport	transport
	system	policy	methodology	scenarios	monitoring

	european	assessment	ctp	mobility	network
	economic	strategic	strategic	models	european
	market	european	project	pricing	information
	cost	level	european	development	freight
	development	development	level	model	making
	analysis	based	modelling	applications	studies

6.4.2.4 Alternative Text Analysis Methods as Points of Reference for Evaluation

Several methods are available for comparing the themes and terms identified by topic modeling against: in the simplest case, one could identify salient terms from the text bodies by computing metrics that represent (weighted) term frequencies, such as tf*idf. Since this thesis is about relational representations of information from texts, I disregard this option, and focus on networks constructed from text data instead: first, for each FP and considered group, I created knowledge networks from the meta-data in the Funding corpus as described in 5.3.2.3. Once the meta-data are organized, e.g. in a database, this approach is about as fast as performing LDA on the texts per group. The entities in the meta-data networks can be considered as a type of ground truth data because they are key words and index terms that were selected by the people who submitted the proposal, and originate from a mixture of pre-defined and self-defined categories that are meant to best represent the gist of a document. Table 123 shows the size of the networks prepared for comparison.

Second, I extracted semantic networks from the text bodies by using the Data to Model (D2M) process as described in section 5.3.2.2.. This process requires a thesaurus. If such a thesaurus has been generated, evaluated and refined, as was done in the previous chapter; extracting knowledge networks this way is also efficient. I reused the refined, auto-generated Funding thesaurus for this purpose; considering all entries as knowledge, which allows for extracting semantic networks instead of meta-networks. Based on my inspection of the semantic networks, I removed a few more overly generic concepts from the thesaurus[23], and regenerated the networks.

Table 123: Size of comparison networks

Data	Groups		Meta-data		D2M+EE	
	Number of members	Number of texts	Nodes	Edges	Nodes	Edges
FP4, group1	21	43	38	169	722	5,521
FP4, group2	16	37	49	246	771	5,278
FP4, group3	13	31	25	111	710	4,980
FP5, group1	147	1,105	209	2,458	3,624	99,960

[23] The removed entries are: 3, 4, including, main, aims, aim.

FP5, group2	85	761	211	2,505	3,047	79,252
FP5, group3	45	534	206	2,364	2,890	60,238
FP6, group1	13	17	66	691	553	3,534
FP6, group2	11	17	84	924	462	2,302
FP6, group3	11	12	60	591	387	1,896

Once these alternative network data are generated, there are several ways for identifying content nodes from them: first, key entity analysis (described in section 5.2.3) can be conducted. This approach has been used in the past for locating content nodes and using them to enhance social network data (described in section 6.1.5.1). To show how the results obtained with topic modeling compare to this common method, I selected this approach for this study. Alternatively, grouping methods could be applied to constructed networks in order to identify groups of structurally similar content nodes. In contrast to key entity analysis of knowledge networks, this approach has not yet been used or validated in this thesis, such that limitations, advantages and typical outcomes of this method in the contexts of this thesis and datasets would be unknown. Also, this approach is not typically used in practical applications. Therefore, I decided to focus on key entity analysis as a point for comparing the outcome per method.

6.4.2.5 Results and Evaluation

There are 120-210 groups per framework program. In order to identify the topics and topic members for the set of texts per groups and comparing these results to knowledge nodes identified with alternative methods, I decided to focus on the three largest groups for FP 4 to FP6. Table 123 shows the size of these groups in terms of members and number of texts. In Table 124 to Table 141, for each group, the following information is presented:

- For topic modeling, the eight most prevalent topics and up to nine topic members[24]. The topics are sorted from left to right by decreasing values of the Dirichlet parameter, which indicates the likelihood of a topic. Green cells indicate the entities that were also found by conducting key player analysis on networks constructed by alternative methods (comparison network).
- For the comparison networks, the ten key entities according to previously introduced network metrics. Green cells indicate terms that are also found among the topic members.

[24] I had planned to retrieve ten members per topic, but in Mallet, the desired number of terms per topic need to set to one more than the number that is retrieved. I only noted this limitation after completing this study.

In all of the results Tables, some terms are abbreviated[25] to accommodate to the real estate on the pages. Each page contains the topic modeling output in the upper table and the results from key entity analysis of both comparison networks in the lower tables. Comparing the results across all three information extraction methods suggests the following:

1. There is a minimal intersection between the key entities from meta-data knowledge networks and topic members from topic modeling. This can be partially explained by the fact that the terms in meta-data are often multi-word combinations of key words, e.g. “sustainable mobility” or “integration of new technology”, while the employed implementation of topic modeling retrieves unigrams only. Considering matches not only on sequence level, but also on the token level might result in higher overlaps.

2. When reading through the members per topic (topic modeling), the terms per topic often mainly related, but it was often hard for me to come up with a fitting label for a topic. This is a well-known limitation of topic modeling. In the past, people have sometimes the strongest word per topic as a topic label, or assigned labels to topics in a qualitative, interpretative fashion. Based on my empirical results, I am proposing an alternative to these strategies: looking at the topics and the key entities from the meta-data network *together*, the highest ranking key entities often seems to be well fitting labels for some of the topics. Here are some examples: in FP6, group 1 (Table 137), the first five topics seem to be about airplanes. For the same data, the key entity from the meta-data networks is “aerospace technology”, which could serve as an appropriate label for these topics. In FP5, group 3 (Table 134), topics 3, 5, and 6-8 seem to be about climate and water. The top entity from the meta-data networks is “environmental protection”. In FP 6, group 3 (Table 141), topics 1-4 and 6 are about tools and products. The corresponding key entity from the meta-data network is “industrial manufacturing”.

3. With topic modeling, while some highly salient terms from the underlying text data occur in multiple topics, most other members appear in one topic. In the meta-data networks and networks extracted from text bodies (in the following referred to as text-based networks), each entity can occur only once per metric, but across metrics, the overlap in key entities is large. Moreover, for both types of comparison networks, the ranking of those entities that occur for

[25] Abbreviations used in table: method. = methodology, develop. = development, tech. = technology, technologies, reg. = regional, interoper. = interoperability, europe. = European, environment. = environmental, info. = information, comm. = communication, transport. = transportation, product. = production, assess. = assessment, apps. = application, applications, manufac. = manufacturing, manufacture, protect. = protection, integrate. = integration, org. = organization, _the_ = _, construct. = construction, intermod. = intermodal, improve. = improvement, monitor. = monitoring, assemble. = assembling

multiple metrics is similar per network construction methods, especially for highly ranked entities.

5. Most of the key entities found in the text-based networks also occur among topic members from multiple topics per text set. This is true for fairly generic terms from the domains of science and research, e.g. “method”, “training” and “integration”, as well as domain specific terms. However, this relationship between text-based networks and topic modeling is asymmetric, i.e. the topic modeling outputs contain many terms that do not occur in the text-based networks. My further in-depth analysis of these terms suggests that they originally also occurred in the auto-generated thesaurus (e.g. “main”, “aims”, “objective”, and “activities”), but got removed from the thesaurus to exclude entities that are overly generic in a dataset and domain. Using the raw, auto-generated thesaurus might have resulted in a higher overlap, but not in more useful network data extracted from the texts. Taking this argument one step further, I suggest that topic modeling; an unsupervised prediction technique, might benefit from the same cleaning techniques that are appropriate for the output of supervised prediction techniques applied to the same data.

6. Discounting for noise terms in topic modeling, the unsupervised prediction approach (topic modeling) and the supervised prediction approach (entity extraction) applied to the same data result in the retrieval of similar terms. This fact partially explains the next finding.

7. In contrast to the key entities from the meta-data networks, the top key entities from the text-based networks would not be useful labels for topics.

8. The key entities in the meta-data and text-based networks are highly similar across the considered metrics per network type. Especially total degree centrality and clique count return similar results, while betweenness centrality provides a complementary set of entities.

Table 124: Topics for FP4, group 1

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
0.25	0.12	0.11	0.08	0.07	0.04	0.03	0.02
policy	transport	transport	transport	projects	intermod.	noise	transport
strategic	europe.	intermod.	data	programme	pre	freight	monitor.
research	project	freight	scenarios	evaluation	transport	track	research
europe.	market	road	mobility	transport	formulas	wagons	centers
method.	objective	project	develop.	project	terminal	traffic	network
project	interoper.	identify	method.	develop.	number	silent	decision
tasks	economic	europe.	pricing	rtd	improve.	europe.	assemble.
ctp	systems	operators	main	framework	policy	low	europe.
level	cost	traffic	socio	options	europe	project	system

Table 125: Key entities for FP4, group 1

Meta-Data				D2M+EE			
Degree centrality	Between. centrality	Eigenvector centrality	Clique count	Degree centrality	Between. centrality	Eigenvector centrality	Clique count
transport	transport	transport	transport	transport	transport	transport	transport
reg._develop.	construct._tech.	reg._develop.	reg._develop.	project	project	europe.	project
construct._tech.	reg._develop.	construct._tech.	construct._tech.	europe.	freight	freight	europe.
safety	policies	safety	safety	freight	projects	infrastructure	method.
policies	sustainable_mobility	policies	policies	model	europe.	intermod.	intermod.
strategic_research	safety	strategic_research	ind._manufac.	intermod.	infrastructure	project	freight
integrate._of_new_tech.	air_transport	integrate._of_new_tech.	economic_aspects	infrastructure	effects	systems	model
tech._transfer	economics_of_transport_systems	tech._transfer	microelectronics	method.	model	monitor.	projects
innovation	quality_of_network	innovation	transports	astra	intermod.	passenger	infrastructure
system_org._and_interoper.	transport_management	system_org._and_interoper.	electronics	projects	criteria	method.	design

Table 126: Topics for FP4, group 2

0.20	0.18	0.14	0.10	0.08	0.06	0.05	0.04
policy	transport	transport	wp	research	dissemination	iea	policy
method.	europe.	urban	develop.	cities	info.	road	scenarios
assess.	public	travel	traffic	europe.	programme	develop.	corridor
define	user	policy	areas	results	project	models	range
project	issues	public	environment.	work	transport	integrated	actions
strategic	potential	uk	decision	case	target	environment.	assess.
projects	users	assess	tools	involve	based	order	countries
ctp	groups	identify	impact	project	aims	lifestyles	economic
task	objective	local	socio	key	impact	project	develop.

Table 127: Key entities for FP4, group 2

Meta-data				D2M+EE			
Degree centrality	Between. centrality	Eigenvector centrality	Clique count	Degree centrality	Between. centrality	Eigenvector centrality	Clique count
transport	transport	transport	transport	transport	transport	transport	transport
reg._develop.	construct._tech.	reg._develop.	safety	strategies	project	strategies	europe.
construct._tech.	policies	construct._tech.	reg._develop.	europe.	europe.	optimal	project
safety	safety	safety	construct._tech.	project	strategies	europe.	strategies
policies	tech._transfer	policies	policies	method.	cities	sustainable	public_transport
tech._transfer	reg._develop.	strategic_research	tech._transfer	cities	eu	rtd	cities
innovation	info._systems	tech._transfer	innovation	public_transport	framework	project	travel
strategic_research	environment._protect.	innovation	environment._protect.	sustainable	public_transport	cities	eu
economic_aspects	industrial_manufac.	integrate._of_new_tech.	economic_aspects	optimal	processes	projects	method.
integrate._of_new_tech.	innovation	economic_aspects	microelectronics	projects	tools	europe	projects

Table 128: Topics for FP4, group 3

0.38	0.15	0.11	0.09	0.08	0.07	0.07	0.06
safety	vts	info.	transport	disc	accident	wp	traffic
project	system	traffic	shipping	demonstration	evacuation	navigation	task
maritime	network	vessel	short	eu	model	gnss	situation
ship	info.	services	sea	training	design	based	develop
transport	project	action	conditions	vii	main	inland	scenarios
assess.	evaluation	vts	test	scenarios	evaluation	info.	comm.
human	comm.	projects	interface	integrated	range	vii	work
operational	processing	users	transport.	control		image	design
related	epto	operators	complete	purposes		radar	obstacles

Table 129: Key entities for FP4, group 3

Meta-Data				D2M+EE			
Degree centrality	Between. centrality	Eigenvector centrality	Clique count	Degree centrality	Between. centrality	Eigenvector centrality	Clique count
transport	transport	transport	transport	vts	vessel	management	vts
reg._develop.	ports_and_logistics)	reg._develop.	reg._develop.	vessel	vts	vessel	transport
construct._tech.	inland_navigation	construct._tech.	construct._tech.	management	transport	services	services
safety	reg._develop.	safety	safety	transport	project	transport	maritime
safety_and_environment_protect._in_maritime_operations	policies	safety_and_environment_protect._in_maritime_operations	policies	eu	services	eu	vessel
efficiency	construct._tech.	efficiency	microelectronics	services	maritime	vts	project
environment._protect.	transports	environment._protect.	industrial_manufac.	project	training	dg	ship
economic_aspects	safety	economic_aspects	electronics	maritime	ship	concept	management
policies	maritime_transport_(shipping	policies	maritime_transport_(shipping	ship	europe.	management_and_info._services	training
maritime_transport_(shipping	telematics_app.s_for_transport	maritime_transport_(shipping	ports_and_logistics)	training	eu	systems	europe.

Table 130: Topics for FP5, group 1

0.76	0.20	0.11	0.09	0.08	0.08	0.07	0.06
project	research	europe.	management	cell	climate	health	product.
develop.	europe.	social	biodiversity	gene	data	clinical	food
develop	network	policy	sustainable	molecular	ocean	disease	treatment
data	info.	economic	land	cells	models	risk	material
based	eu	eu	europe	expression	carbon	control	waste
results	international	countries	environment.	genes	chemical	europe	products
environment.	workshops	public	water	disease	europe.	food	mesh
provide	activities	policies	forest	protein	time	treatment	water
quality	scientific	develop.	conservation	mechanisms	model	diseases	gauge

Table 131: Key entities for FP5, group 1

Meta-Data				D2M+EE			
Degree centrality	**Between. centrality**	**Eigenvector centrality**	**Clique count**	**Degree centrality**	**Between. centrality**	**Eigenvector centrality**	**Clique count**
environment._protect.	training	environment._protect.	economic_aspects	project	project	management	project
life_sciences	policies	life_sciences	scientific_research	europe.	europe.	fisheries	europe.
economic_aspects	environment._protect.	economic_aspects	environment._protect.	management	europe	europe.	management
scientific_research	education	fisheries	social_aspects	fish	analysis	project	fish
fisheries	renewable_sources_of_energy	resources_of_sea	policies	fisheries	study	fish	studies
resources_of_sea	tech._transfer	agriculture	regulations	analysis	network	aquaculture	analysis
health	social_aspects	food	legislation	eu	eu	sustainable	models
medicine	reg._develop.	resources_of_sea_fisheries	renewable_sources_of_energy	species	studies	species	model
agriculture	scientific_research	key_action_sustainable_agriculture	meteorology	models	model	eu	fisheries
policies	transport	fisheries_and_forestry	life_sciences	methods	systems	marine	eu

Table 132: Topics for FP5, group 2

0.84	0.45	0.25	0.18	0.11	0.08	0.07	0.07
project	project	europe.	policy	materials	energy	system	system
develop.	models	network	environment.	material	power	fuel	based
tech.	data	research	economic	components	system	energy	monitor.
product.	model	projects	policies	high	renewable	power	tool
process	results	knowledge	energy	process	pv	heat	optical
high	tools	eu	impacts	parts	systems	cell	control
cost	analysis	activities	sustainable	coatings	solar	hybrid	machine
systems	test	info.	develop.	manufac.	market	cooling	software
develop	based	countries	framework	composite	integrate.	efficiency	refurbishment

Table 133: Key entities for FP5, group 2

Meta-data				D2M+EE			
Degree centrality	Between. centrality	Eigenvector centrality	Clique count	Degree centrality	Between. centrality	Eigenvector centrality	Clique count
economic_aspects	standards	economic_aspects	economic_aspects	project	project	project	project
environment._protect.	evaluation	environment._protect.	environment._protect.	europe.	europe.	energy	systems
scientific_research	environment._protect.	innovation	scientific_research	energy	systems	systems	design
industrial_manufac.	social_aspects	industrial_manufac.	social_aspects	systems	energy	design	energy
renewable_sources_of_energy	renewable_sources_of_energy	safety	policies	design	europe	europe.	europe.
energy_saving	policies	tech._transfer	regulations	tools	eu	tools	performance
social_aspects	reg._develop.	materials_tech.	legislation	models	tools	tech.	models
tech._transfer	fisheries	energy_saving	energy_saving	analysis	models	advanced	tech.
innovation	tech._transfer	renewable_sources_of_energy	renewable_sources_of_energy	fuel	analysis	analysis	advanced
safety	other_energy_topics	key_action_innovative_products	other_energy_topics	tech.	app.s	fuel	tools

Table 134: Topics for FP5, group 3

0.80	0.16	0.11	0.09	0.08	0.07	0.07	0.07
project	research	climate	policy	coastal	ozone	water	materials
provide	europe.	models	urban	marine	chemical	ecosystems	tech.
based	network	model	economic	mediterranean	atmospheric	management	industrial
results	social	data	decision	sea	climate	biodiversity	high
develop.	access	ocean	develop.	water	impact	community	process
develop	info.	sea	air	ecosystem	aerosol	natural	product.
systems	europe	variability	mountain	product.	emissions	europe	efficiency
developed	activities	system	policies	species	atmosphere	species	cost
info.	national	atmospheric	eu	waters	processes	fishing	develop.

Table 135: Key entities for FP5, group 3

Meta-data				D2M+EE			
Degree centrality	Between. centrality	Eigenvector centrality	Clique count	Degree centrality	Between. centrality	Eigenvector centrality	Clique count
environment._protect.	environment._protect.	environment._protect.	scientific_research	project	project	project	project
economic_aspects	policies	fisheries	economic_aspects	europe.	europe.	europe.	europe.
scientific_research	social_aspects	resources_of_sea	environment._protect.	models	europe	models	model
fisheries	scientific_research	forecasting	social_aspects	model	analysis	model	models
resources_of_sea	standards	mathematics_statistics	policies	analysis	models	expected	analysis
social_aspects	education_and_training	meteorology	regulations	systems	model	modeling	systems
life_sciences	industrial_manufac.	measurement_methods	legislation	europe	studies	approach	europe
meteorology	info._processing	climate_and_biodiversity	meteorology	management	novel	impacts	modeling
measurement_methods	renewable_sources_of_energy	key_action_global_change	renewable_sources_of_energy	ozone	systems	management	understanding
forecasting	reg._develop.	economic_aspects	life_sciences	studies	study	systems	studies

Table 136: Topics for FP6, group 1

0.26	0.11	0.07	0.07	0.06	0.04	0.04	0.03
engine	aircraft	tbc	turbine	noise	industry	project	process
low	concepts	control	engine	broadband	automotive	research	equipment
noise	capabilities	provide	cfd	methods	innovative	field	significant
aircraft	future	key	aero	prediction	tech.	europe	supply
vital	integrate.	tech.	aggressive	research	range	goals	breakthrough
tech.	assess.	aero	technical	fan	methods		
engines	impact		high	understanding	low		
fan			environment	programmes	provide		
weight			goal	universities			

Table 137: Key entities for FP6, group 1

Meta-data				D2M+EE			
Degree centrality	**Between. centrality**	**Eigenvector centrality**	**Clique count**	**Degree centrality**	**Between. centrality**	**Eigenvector centrality**	**Clique count**
aerospace_tech.	propulsion	aerospace_tech.	aerospace_tech.	noise	project	noise	noise
measurement_methods	aerospace_tech.	forecasting	forecasting	low	aircraft	low	engine
mathematics_statistics	evaluation	mathematics_statistics	mathematics_statistics	engine	noise	fan	aircraft
forecasting	environment._protect.	measurement_methods	measurement_methods	aircraft	engine	engine	methods
innovation	cooperation	tech._transfer	industrial_manufac.	fan	europe.	broadband	design
tech._transfer	systems_approach_to_future_efficient	policies	tech._transfer	tech.	advanced	aircraft	project
policies	industrial_manufac.	innovation	policies	project	methods	turbo	industry
economic_aspects	social_aspects	social_aspects	innovation	europe.	industry	concepts	advanced
social_aspects	coordination	evaluation	economic_aspects	methods	improved	tech.	tech.
evaluation	policies	economic_aspects	environment._protect.	design	novel	weight	low

Table 138: Topics for FP6, group 2

0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01
europe.	control	track	risk	noise	industrial	samco	bridge
research	vibration	methods	building	vehicles	system	international	high
transport	adaptive	network	develop.	measures	systems	structural	market
integrated	impact	project	assess.	impact	assess.	field	modtrain
system	design	countries	tech.	approaches	monitor.	thematic	product
tech.	landing		design	control	objective		tech.
systems	shock		activities		safety		
objective	structural				risk		
services	full					integrated	

Table 139: Key entities for FP6, group 2

Meta-data				**D2M+EE**			
Degree centrality	**Between. centrality**	**Eigenvector centrality**	**Clique count**	**Degree centrality**	**Between. centrality**	**Eigenvector centrality**	**Clique count**
innovation	industrial_manufac.	tech._transfer	tech._transfer	design	design	network	design
tech._transfer	construct._tech.	innovation	innovation	network	systems	operators	components
policies	evaluation	policies	scientific_research	structural	bearings	eight	energy
environment._protect.	transport	environment._protect.	policies	methods	integrated	project	systems
scientific_research	environment._protect.	energy_saving	industrial_manufac.	systems	europe.	function	building
energy_saving	safety	renewable_sources_of_energy	measurement_methods	europe.	advanced	validation	structural
renewable_sources_of_energy	measurement_methods	fossil_fuels	evaluation	infrastructure	project	europe	bearings
fossil_fuels	media	other_energy_topics	forecasting	solutions	performance	infrastructure	integrated
other_energy_topics	policies	scientific_research	environment._protect.	project	road_transport	railways	europe.
industrial_manufac.	tech._transfer	fisheries	energy_saving	integrated	energy	db	adaptive

Table 140: Topics for FP6, group 3

0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01
tooling	micro	particles	industrial	kmm	product.	coated	tactile
adjustable	products	products	forging	integrate.	demands	sheet	neural
manufac.	manufac.	project	virtual	training	manufac.	polymer	virtual
tech.	mass	develop	knowledge	micro	integrate.	develop	sensors
forming	tech.	objective	materials	europe.	systems	based	
innovative	systems	integrate.	processes		products	products	
project	integrated	micro	create				
	training		related				
	develop		integrate.				

Table 141: Key entities for FP6, group 3

Meta-data				D2M+EE			
Degree centrality	**Between. centrality**	**Eigenvector centrality**	**Clique count**	**Degree centr.**	**Between. centr.**	**Eigenvector centr.**	**Clique count**
industrial_manufac.	industrial_manufac.	industrial_manufac.	industrial_manufac.	tooling	design	tooling	tooling
tech._transfer	tech._transfer	innovation	aerospace_tech.	virtual	products	materials	design
innovation	biotech.	tech._transfer	forecasting	design	micro	virtual	products
innovation_tech._transfer	new_and_user-friendly_product._equipment_and_tech.	innovation_tech._transfer	mathematics_statistics	materials	tech.	simulations	processes
materials_tech.	aerospace_tech.	materials_tech.	measurement_methods	micro	europe.	processes	key
and_their_incorporation_into_factory_of_future	cooperation	cooperation	tech._transfer	processes	tooling	led	tools
coordination	and_their_incorporation_into_factory_of_future	and_their_incorporation_into_factory_of_future	innovation	products	project	design	led
new_and_user-friendly_product._equipment_and_tech.	based_on_nanotech._and_new_materials	coordination	scientific_research	europe.	advanced	key	materials
cooperation	measurement_methods	new_and_user-friendly_product._equipment_and_tech.	innovation_tech._transfer	led	processing	testing	advanced
aerospace_tech.	coordination	biotech.	materials_tech.	knowledge	led	products	europe.

6.4.3 Application Context III: Enron Corpus

6.4.3.1 Social Network Data

For the social networks, I re-used the communication networks that I had constructed from the Enron email headers as described in section 5.3.2.3. The communication networks are weighted, directed graphs. For information about the considered time periods and sizes of the networks see Table 114.

6.4.3.2 Grouping of Social Network Data

First, isolates were removed from the communication networks, since they would only form groups of their own or with other isolates. Furthermore, I dropped loops, which happen if people copy or blindcopy themselves on an email. I did not remove pendants, which for these data are people who only receive emails, but did not send an email to anybody in the considered sample. However, in the context of covert networks, people who only receive information have shown to be highly relevant: when planning and executing illicit activities, the need to conceal is higher than the need to coordinate (Baker & Faulkner, 1993). Consequently, people tend to keep their communication volumes low (Klerks, 2001).

The social networks from the Enron data are denser than the Funding networks. This is partially due to the chosen data construction mechanism: the Funding data are star network structures around PIs, while for Enron, all emails sent or received by the people in the CASOS Enron database are represented as links.

In contrast to the Funding data, for the Enron networks, groups based on CONCOR were not mainly based on the number of emails that people had sent or received. However, and as already observed for the Funding data, the members within CONCOR groups typically do not share direct connections, but were spread across the network. Therefore, the same argument as made before applies here as well, namely that enforcing shared content onto these group members seems to be an inappropriate strategy as it may result in false positive links.

Due to the comparatively high network density, the Girvan-Newman algorithm found less distinct groups in the Enron communication networks than in the Funding collaboration networks. In fact, without any network post-processing of the Enron networks, the vast majority of nodes get places into one group and also into one component. In order to explore whether removing low-weight nodes can help to mitigate this issue, i.e. breaking apart clusters, I identified meaningful cut-off values for links to disregard for grouping as follows: I inspected the in-degree and out-degree distribution of the networks (Figure 15, Figure 16); realizing that they do not follow the classic power law distribution. This means that it is not the case that most people have a low email volume, especially not for emails received. Since this observation is a counterargument to the previous point that people involved in illicit activities keep their communication volumes low, it further supports the previously emphasized fact that much of the conversation and many of the people in Enron had nothing to do with any illicit activities.

Figure 15: Distribution of emails sent

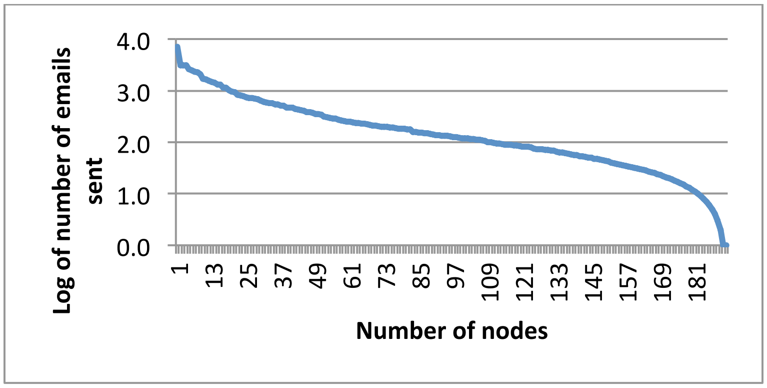
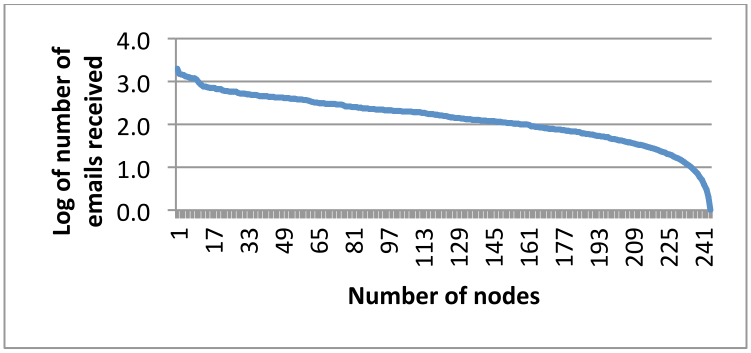
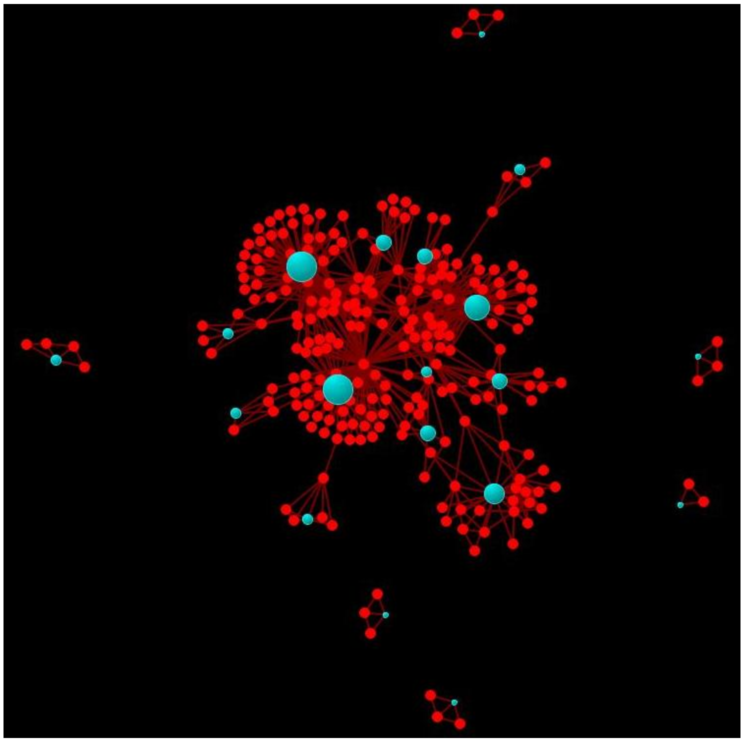


Figure 16: Distribution of emails received



Further inspecting the link frequency distributions, I decided to drop links with a frequency of less than 16. Applying Girvan-Newman again did resulted in multiple groups, but visually inspecting them in ORA suggested that the larger groups still had sub-structures that Girvan-Newman did not pick up on yet. Therefore, for each of the three networks, I increased the number of Girvan-Newman groups one by one, visually inspected the resulting partitioning, and identified the most appropriate number of groups through this visual analytics procedure. Figure 17 shows an example of this process, where the final groups for time period 1 are displayed (groups are indicated by the green circle that holds the group members together). Next, I passed this number as a parameter to the Girvan-Newman algorithm. Comparing the resulting groups for the visual and purely algorithm-based Girvan-Newman grouping showed that as one would expect, the retrieved groups coincided.

Figure 17: Example for Girvan-Newman groups in Enron, time period 1



The number and size of groups per time period considered is shown in Table 142. Overall, groups in these data center on people who sent one or more emails to many other people. While these groups can also be retrieved by extracting the ego-network of key entities that score highest on node centrality metrics, these small, disjoint groups would be missed with this alternative approach.

Table 142: Number and size of networks and groups

Data	Raw			Groups			Number of groups					
	Nodes	Edges	Emails	Nodes	Edges	Modularity	Count	Min	Max	Average	Std Dev	10+ nodes
Period 1	448	3,092	6,901	238	498	15.3	19	2	48	12.5	14.1	7
Period 2	433	2,295	3,711	151	234	24.4	11	2	66	13.7	20.6	3
Period 3	435	4,721	11,042	322	1,099	22.4	10	8	124	32.1	35.3	8

6.4.3.3 Identify Content Nodes per Group via Topic Modeling

For each group, I retrieved the emails sent among members of the groups. This design decision deviates from the Funding data, where I also considered proposals that PIs had co-authored with people outside their groups since the group might still benefit from this expertise. However, email data is more private than scientific information, and it is not a given that a group has access to the knowledge that some group members shares with somebody outside the group.

For topic modeling in Mallet, I explored different numbers of topics again[26]. For all other parameters, I used the same settings as for the Funding data. Based on my screening and comparison of the results, I decided to generate the numbers of topics as shown in Table 143. One reason for why the number of potentially useful topics does not linearly increase with the number of texts is that the same email might occur in multiple people's inboxes, e.g. when somebody sent or forwarded an email to multiple recipients.

6.4.3.4 Alternative Text Analysis Methods as Point of Reference for Evaluation

For the Enron data, we have no meta-data available that can serve as a point of comparison. Therefore, I only extracted networks from the email bodies per group and time period as follows: I re-used the refined, auto-generated Enron thesaurus as part of the D2M text coding process. Since we only need knowledge node here and topic modeling does not differentiate between different node classes either, I converted all but the attribute entries in the thesaurus to be associated with the knowledge class. Also, I removed a few more numerical entries (all numbers from 1 to 150) that should have been classified as attributes. The resulting thesaurus has 6,227 entries. Table 143 shows the number of nodes in the groups and comparison networks. Both topic modeling and key entity analysis are based on the exact same text data.

Table 143: Size of groups and comparison networks

Data		Social Network		Topics	D2M+EE	
Time Period	Group	Members	Texts		Nodes	Edges
1	1	48	189	15	612	2,090
1	2	44	133	15	581	1,430
1	3	33	442	20	1,388	9,786
2	1	66	240	15	867	2,626
2	2	33	1,212	25	4,068	44,370
2	3	28	489	20	1,151	5,622
3	1	124	1,931	25	2,025	14,026
3	2	51	418	20	1,146	6,052
3	3	37	437	20	1,101	5,176

6.4.3.5 Results and Evaluation

To stay consistent with the approach to data analysis and evaluation used for the Funding data, I analyze the top three groups per time period again. The same network metrics as used for the Funding comparison networks are employed again for the Enron text-based networks. However, in order to provide some additional information about the relationship between topic modeling

[26] This time, I requested the top eleven terms in order to get the top ten terms.

and key entities from text-based networks, I use a different way of presenting the results: Table 144 to Table 152 each show the outcome of both methods; containing the following:

- The first block are the terms identified by both topic modeling and key entity analysis of the text-based networks. The comparison is based on the top ten topics from topic modeling and the top ten key entities from the text-based networks.
- The second block lists the entities found only via key entity analysis of the text networks.
- The third block shows the topics and members not found in the comparison network.

The following results from the Funding application context can be confirmed with the results from this application context:

1. Most of the key entities from the text-based networks are also retrieved with topic modeling. This is true for generic terms from the domain and dataset, e.g. "Enron" and instances of the time entity class, as well as specific terms. This relationship between text-based networks and topic modeling is asymmetric: the topic modeling outputs contain many terms that do not occur in the text-based networks, but this might be mainly due to the limited number of key-entities retrieved.

2. Further analyzing the terms found with topic modeling, but not key entities analysis, shows that many of these terms were originally in the auto-generated, refined thesaurus, but eliminated as part of the thesaurus cleaning process, e.g. "pmto" and "amto". I had removed these entities from the thesaurus to exclude overly generic terms given the dataset and domain. Using the raw thesaurus might have resulted in a higher overlap, but not in more useful networks.

3. After disregarding noise terms from topic modeling, the unsupervised and the supervised prediction methods result in the retrieval of similar terms, which is limited by the number of key entities from text networks considered for this comparison.

4. The top key entities from the text-based networks would not be useful labels for topics.

There are additional findings that apply to the Enron data, but not to the Funding data:

5. On a qualitative level, both information extraction methods return less meaningful results than with the Funding data. For example, entities consistently ranked high with both methods include "Enron", "energy", and time terms. This might be because the email data are nosier than the funding proposal descriptions. For example, in forwarded messages, the email bodies contain time stamps and names of other people, which are reflected in both sets of results. However, this finding suggests again an agreement between the supervised and unsupervised prediction models.

6. The topics seem harder to distinct than for the Funding data, i.e. the similar gist of information seems to be suggested by multiple topics per run. This could be due to the data itself or due to

the high similarity among the documents per group, which could happen for instance if multiple people have the same or similar email in their inbox. The same effect might have been observed for the Sudan data, where the texts per group highly overlapped.

Table 144: Topics and Key Entities, Time period 1, Group 1

	Topic										Network Metrics			
Entity	1	2	3	4	5	6	7	8	9	10	Degree Centrality	Between. Centrality	Eigenvec. Centrality	Clique Count
	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01				
enron			x	x				x			x	x	x	x
june		x								x	x	x		x
david						x					x		x	x
energy								x			x	x		x
doug			x								x		x	
john					x							x		x
tom						x					x		x	
gas								x				x		x
steve									x		x		x	
entergy		x								x				x
hernandez		x											x	
unit					x							x		
kayne							x							x
ees								x				x		
ferc										x		x		
sent											x	x		x
miller											x		x	
please												x		x
chad													x	
mike													x	
robert											x			
watts													x	
day	x	x												
ercot	x				x									
market	x								x					
baughman	x													
don	x													
group	x													
hourly	x													
notes	x													
questions	x													
real	x													
subject		x		x										
bill		x												
coulter		x												
juan		x												
lloyd		x												
request		x												

ect			x	x					x		
hou			x	x					x		
pm			x	x							
corp			x								
enronxgate			x								
joe			x								
larry			x								
na			x								
power				x				x			
smith				x					x		
forwarded				x							
pmto				x							
tecoenergy				x							
deal					x						
enpower					x						
lcra					x						
list					x						
message					x						
mw					x						
org					x						
mdea						x	x				
reagan						x	x				
bogey						x					
chose						x					
commercial						x					
contract						x					
mann						x					
time						x					
customer							x				
data							x				
draft							x				
epmi							x				
jeff							x				
load							x				
section							x				
million								x			
risk								x			
rogers								x			
trading								x			
wholesale								x			
access									x		
bid									x		
model									x		
options									x		
tamara									x		
agreement										x	
amrn										x	

er										x				
filing										x				
Interconnect.										x				
mapp										x				
settlement										x				

Table 145: Topics and Key Entities, Time period 1, Group 2

	Topics										Network Metrics			
Entity	1	2	3	4	5	6	7	8	9	10	Degree Centr.	Between. Centr.	Eigenv. Centr.	Clique Count
	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01				
enron			x	x	x	x			x		x	x	x	x
energy						x					x	x	x	x
june	x		x					x			x	x		x
chris	x				x				x		x	x		x
wednesday	x										x		x	x
analyst						x					x		x	x
gas								x			x	x		x
john							x		x				x	x
firm						x							x	
transco								x				x		
gov									x			x		
sent											x		x	x
thursday											x		x	
212											x			
bob												x		
capacity												x		
doug													x	
joseph												x		
street													x	
plants														x
original	x		x		x					x				
message	x				x					x				
amto	x													
chrissent	x													
dorland	x													
items	x													
jpg	x													
pm		x								x				
add		x												
click		x												
excel		x												
exotica		x												
library		x												
meeting		x												
option		x												
options		x												
time		x												
email			x							x				

book	x								
canadian	x								
deal	x								
matt	x								
sold	x								
volume	x								
year		x			x				
corp		x					x		
subject		x						x	
data		x							
days		x							
load		x							
mw		x							
pmto		x							
smith		x							
ect			x	x			x		
beer			x						
germany			x						
nyiso			x						
symptom			x						
ve			x						
earnings				x					
news				x					
ows				x					
revenue				x					
tor				x					
don					x				
lagrasta					x				
list					x				
mark					x				
model					x				
notes					x				
people					x				
trading					x				
aep						x			
bcf						x			
gri						x			
high						x			
paul						x			
rates						x			
supply						x			
mail							x	x	
ca							x		
offer							x		
org							x		
fw								x	
king								x	
ny								x	
read								x	

Table 146: Topics and Key Entities, Time period 1, Group 3

Entity	1	2	3	4	5	6	7	8	9	10	Degree Centr.	Between. Centr.	Eigenv. Centr.	Clique Count
	0.10	0.08	0.06	0.06	0.06	0.05	0.04	0.03	0.02	0.02				
enron	x			x			x	x	x		x	x	x	x
jeff	x						x	x			x	x	x	x
california		x					x				x	x		x
davis		x									x	x		x
energy		x									x	x		x
mara	x										x		x	
susan	x										x		x	
ferc						x						x		x
dasovich								x			x		x	
john			x											x
time			x									x		
governor					x			x						x
bill												x		x
ees											x		x	
415											x			
david													x	
gov												x		
government														x
james													x	
richard													x	
sent												x		
steffes													x	
pm	x								x					
subject	x								x					
corp	x													
fax	x													
forwarded	x													
na	x													
market		x				x			x					
power		x				x								
prices		x				x								
electricity		x												
generators		x												
state		x												
utilities		x												
alan			x											
capacity			x											
day			x											
gas			x											
message			x											
original			x											
pg			x											
sce			x											
contracts				x	x									
customers				x						x				

core	x							
dwr	x							
hertzberg	x							
noncore	x							
past	x							
rate	x							
rates	x							
group		x		x				
mail		x					x	
june		x						
bankruptcy		x						
financial		x						
mou		x						
plan		x						
qfs		x						
bush			x					
cap			x					
caps			x					
commission			x					
order			x					
price			x					
call				x				
folks				x				
hoffman				x				
meeting				x				
solution				x				
week				x				
enronxgate					x			
govenar					x			
investments					x			
michael					x			
million					x			
news					x			
ca						x		
caiso						x		
confidential						x		
iso						x		
jeanne						x		
participants						x		
access							x	
direct							x	
edison							x	
manuel							x	
org							x	
puc							x	
tracy							x	
users							x	

Table 147: Topics and Key Entities, Time period 2, Group 1

	Topics										Network Metrics			
Entity	1	2	3	4	5	6	7	8	9	10	Degree	Between.	Eigenv.	Clique
	0.07	0.07	0.06	0.03	0.03	0.03	0.02	0.02	0.02	0.01	Centr.	Centr.	Centr.	Count
enron			x				x	x			x	x	x	x
august	x										x		x	x
september										x	x		x	x
tuesday	x										x		x	
chris	x										x			x
ercot		x	x	x								x		x
john			x			x					x	x		
time	x											x		
wednesday							x						x	
sent											x	x	x	x
monday											x		x	x
thursday											x		x	x
mike											x		x	x
friday													x	x
october												x		
november												x		
energy												x		
gas												x		
please												x		
message	x		x					x						
original	x		x					x						
pmto	x		x											
amto	x													
cowan	x													
dorland	x													
mw		x			x									
frontera		x												
hour		x												
jmf		x												
oom		x												
plan		x												
plant		x												
price		x												
resource		x												
forney			x											
joe			x											
mark			x											
subject			x											
load				x	x									
discuss				x										
list				x										
north				x										
options				x										
products				x										

trades	x							
trading	x							
zonal	x							
existing		x						
information		x						
main		x						
peak		x						
physical		x						
place		x						
power		x						
transmission		x						
scott			x	x				
call			x			x		
don			x			x		
group			x			x		
attached			x					
louise			x					
michael			x					
questions			x					
robert			x					
doc				x				
email				x				
ensr				x				
kevin				x				
received				x				
side				x				
week				x				
mail					x	x		
annulled					x			
corp					x			
duplicate					x			
intended					x			
pjm					x			
recipient					x			
communication						x		
day						x		
holiday						x		
national						x		
number						x		
work						x		
corporation							x	
cows							x	
due							x	
eps							x	
event							x	
exc							x	
major							x	
markets							x	
stocks							x	

Table 148: Topics and Key Entities, Time period 2, Group 2

	Topics										Network metrics			
Entity	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	Degree Centr.	Between. Centr.	Eigenv. Centr.	Clique Count
	0.12	0.09	0.08	0.06	0.05	0.04	0.04	0.03	0.03	0.03				
enron	x	x	x		x			x			x	x	x	x
september	x		x				x				x		x	x
energy					x						x	x		x
august	x	x									x		x	
jeff	x												x	x
ferc		x										x		x
dasovich	x												x	
james		x											x	
company					x						x			
gas					x						x			
time					x							x		
bill						x								x
sent											x	x	x	x
california											x	x		x
ees											x		x	
john												x		x
monday												x		
wednesday													x	
friday													x	
dynegy												x		
electric											x			
scheduling												x		
week														x
message	x	x	x				x							
original	x	x	x				x							
pmto	x	x												
susan	x									x				
mara	x													
christi		x												
jim		x												
steffes		x												
amto			x											
herndon			x											
kevin			x											
presto			x											
risk			x											
rogers			x											
customers				x				x						
july				x										
access				x										
contracts				x										
da				x										
date				x										
decision				x										
direct				x										

dwr				x							
puc				x							
business					x						
home					x						
services					x						
tax					x						
williams					x						
assembly						x					
committee						x					
davis						x					
edison						x					
governor						x					
legislature						x					
provisions						x					
senate						x					
session						x					
draft							x		x		
mail							x			x	
arem							x				
dan							x				
douglass							x				
energyattorney							x				
mailto							x				
cpuc								x	x		
approach								x			
credit								x			
credits								x			
ctc								x			
pg								x			
px								x			
sce								x			
authority									x		
commission									x		
court									x		
federal									x		
filing									x		
issues									x		
petition									x		
rehearing									x		
donna										x	
frank										x	
group										x	
linda										x	
paul										x	
robert										x	
steve										x	
work										x	

Table 149: Topics and Key Entities, Time period 2, Group 3

	Topics										Network Metrics			
	1	2	3	4	5	6	7	8	9	10	Degree Centr.	Between. Centr.	Eigenv. Centr.	Clique Count
Entity	0.24	0.08	0.08	0.08	0.04	0.04	0.03	0.03	0.03	0.02				
thursday	x										x	x	x	x
enron					x	x		x	x		x	x	x	x
august	x								x		x		x	x
september	x										x		x	x
wednesday	x										x		x	x
gas		x	x								x	x		x
energy			x									x		x
kim			x										x	
sent											x	x	x	x
tuesday											x		x	x
please												x		x
713											x			
monday													x	
friday													x	
company												x		
david												x		
houston												x		
nymex											x			
week												x		
message	x			x			x	x						
original	x			x			x							
pmto	x						x							
amto	x													
fw	x													
subject	x													
mark		x		x		x								
barry		x												
bt		x												
deal		x												
group		x												
storage		x												
tycholiz		x												
west		x												
year		x												
credit			x			x								
dwr			x											
natural			x											
power			x											
price			x											

trade			x								
trading			x								
cheryl				x							
eol				x							
greenberg				x							
jones				x							
legal				x							
tana				x							
taylor				x							
attached					x	x					
corp					x			x			
fax					x					x	
america					x						
cook					x						
cordially					x						
mary					x						
north					x						
texas					x						
agreement						x					
comments						x					
dth						x					
isda						x					
nda						x					
questions						x					
frank							x		x		
allen							x				
grigsby							x				
jay							x				
mike							x				
scott							x				
tori							x				
contract								x			
heard								x			
intended								x			
mail								x			
mailto								x			
marie								x			
recipient								x			
greg									x		
kaminski									x		
predict									x		
stanford									x		
trip									x		

vince									x					
wolak									x					
ahouston										x				
sara										x				
securities										x				
shackleton										x				
shackletonenron										x				
smith										x				
street										x				
suchdev										x				
tx										x				

Table 150: Topics and Key Entities, Time period 3, Group 1

	Topics										Network metrics			
Entity	1	2	3	4	5	6	7	8	9	10	Degree Centr.	Between. Centr.	Eigenv. Centr.	Clique Count
	0.05	0.04	0.04	0.04	0.04	0.03	0.03	0.02	0.02	0.02				
enron	x			x	x			x	x		x	x	x	x
november	x		x		x						x		x	x
october	x										x		x	x
monday				x							x	x	x	
john							x	x			x		x	x
tuesday	x										x		x	
mike			x					x			x			x
gas									x	x	x			x
wednesday	x												x	
ercot						x		x				x		
energy						x			x					x
time						x						x		
david							x					x		
sent											x	x	x	x
please											x	x		x
friday													x	x
august												x		
thursday													x	
doug												x		
smith												x		
message	x		x	x	x		x	x						
original	x		x	x	x		x	x						
pmto	x		x		x		x							
amto	x													
fw	x													
deal		x				x								
back		x												
book		x												

kam	x									
list	x									
make	x									
netco	x									
process	x									
start	x									
week	x									
chris		x								
desk		x								
dorland		x								
grigsby		x								
phillip		x								
day			x							
don			x							
group			x							
pjm			x							
pm			x							
work			x							
meeting				x		x	x			
curves				x						
data				x						
file				x						
subject				x						
power					x			x		
load					x					
market					x					
mw					x					
price					x					
sell					x					
integration						x				
kitchen						x				
louise						x				
webb						x				
ees							x			
greg							x			
mark							x			
company								x		
credit								x		
mail								x		
marketing								x		
trading								x		
transactions								x		
business									x	

daily										x				
keystone										x				
mexican										x				
operations										x				
socal										x				
storage										x				
units										x				
weather										x				

Table 151: Topics and Key Entities, Time period 3, Group 2

	Topics										Network Metrics			
Entity	1	2	3	4	5	6	7	8	9	10	Degree	Between.	Eigenv.	Clique
	0.08	0.05	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	Centr.	Centr.	Centr.	Count
california				x							x	x	x	x
enron			x				x			x		x		x
deals				x	x						x		x	
energy									x			x		x
november	x				x									x
october	x						x							x
epmi				x									x	
palo				x							x			
john								x				x		
epmi_short_term											x		x	x
southwest											x		x	x
daily											x		x	
epmi_long_term											x		x	
mwh											x		x	
northwest											x		x	
sent											x			x
monday														x
bill												x		
company												x		
dynegy														x
eol													x	
ferc												x		
gas												x		
issue												x		
jim												x		
filename	x	x												
message	x				x		x		x					
thursday	x				x									
original	x						x							
amto	x													
holden	x													
salisbury	x													
tim	x													
deal		x			x									

diana	x								
load	x								
mw	x								
pge	x								
questions	x								
scheduling	x								
system	x								
transmission	x								
power		x			x				
price		x			x				
west		x			x				
america		x							
bruce		x							
comments		x							
conference		x							
north		x							
report		x							
day			x				x		
long			x						
short			x						
sp			x						
term			x						
total			x						
group				x			x		
build				x					
cara				x					
dart				x					
fran				x					
market					x			x	
cash					x				
current					x				
desk					x				
marketing					x				
prices					x				
receive					x				
fw						x			
http						x			
mailto						x			
mm						x			
pmto						x			
tag						x			
gov							x		
information							x		
intended							x		
mail							x		
provide							x		
work							x		
wscc							x		
alan								x	
caiso								x	

iso									x					
notice									x					
rto									x					
time									x					
week									x					
christian										x				
contract										x				
customer										x				
hall										x				
kind										x				
mike										x				
office										x				
year										x				
yoder										x				

Table 152: Topics and Key Entities, Time period 3, Group 3

	Topics										Network metrics			
Entity	1	2	3	4	5	6	7	8	9	10	Degree	Between.	Eigenv.	Clique
	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	Centr.	Centr.	Centr.	Count
enron						x			x		x	x	x	x
gas	x										x	x		x
november				x		x	x				x		x	x
david				x							x		x	x
october						x					x		x	x
monday							x				x		x	
thursday				x		x							x	
week					x							x		
team								x			x			
sent											x	x	x	x
john											x	x		x
tuesday												x	x	
company											x			x
energy												x		x
please												x		x
august												x		
friday													x	
choate													x	
new_york												x		
message	x	x		x		x	x		x					
smith	x	x												
original	x			x		x	x		x					
scott	x							x						
bateseast	x													
judy	x													
kimberly	x													
mckay	x													
vladi	x													

call	x			x				
america	x							
debra	x							
eb	x							
fax	x							
legal	x							
meeting	x							
street	x							
balance		x						
book		x						
contract		x						
cuilla		x						
curve		x						
egan		x						
leach		x						
martin		x						
point		x						
robert		x						
pmto			x		x	x		
subject			x		x	x		
december			x					
baumbach			x					
love			x					
asked				x				
called				x				
deal				x				
demand				x				
list				x				
time				x				
today				x				
told				x				
doc					x			
recipient					x			
amto						x		
fw						x		
http						x		
mail						x		
commercial							x	
desk							x	
directly							x	
logistics							x	
mike							x	
neal							x	
report							x	
shively							x	
agreement								x
comments								x

language	x		
master	x		
nicor	x		
party	x		
review	x		
added		x	
comwww		x	
deleted		x	
folder		x	
inbox		x	
item		x	
items		x	
offline		x	
synchronizing		x	
updated		x	

6.5 Conclusions

Based on leveraging the concept of social groups, a computational and interdisciplinary methodology for jointly considering text data and network data was developed, operationalized and tested in real-world application scenarios. The resulting methodology facilitates the enhancement of social network data with content nodes and fixes the main limitation with this approach, namely the arbitrary identification of content nodes and which agents these nodes are linked to. The proposed methodology scales up to large corpora. At the same, the methodology allows for gaining an in-depth understanding of the content that groups of structurally coherent agents are exposed to directly or within a few degrees of separation in their social network. However, further work would is to fully automate this process. The next section (6.6) suggests some strategies towards that goal.

The methods review in this chapter has led to the following conclusions. First, extracting content nodes from groups of *structurally equivalent* agents is an appropriate strategy for enabling the *comparison* of the content that these agents produce, perceive or disseminate. Also, these equivalence classes can represent a variety of social roles and positions that network members can occupy. These roles include classic network power roles that are defined over node centrality metrics, other structurally defined roles, such as formal and informal leaders, and roles defined over behavioral signatures, such as homophily. Second, extracting content nodes from groups of *structurally coherent* agents is an appropriate strategy for enabling the *enhancement* of social network data with content nodes. Since this enhancement process is the primary goal with this chapter, the second strategy was selected for further work. The developed approach provides an answer to the first research question addressed in this chapter, which was:

- How can the method of enhancing social networks with content nodes be advanced such that the arbitrariness of adding content nodes to social networks is mitigated?

Two more methodological research questions were addressed in this chapter, which aimed at comparing the outcome of the proposed method to alternative methods. These questions were:

- How do key entities identified by applying certain prediction models trained with supervised learning compare to entities identified via topic modeling?
- How does topic modeling compare to alternative information or relation extraction methods in terms of identified terms and themes?

The following answers to these questions were found by operationalizing the proposed methodology and testing its application to various datasets:

First, even though the overlap between key entities from meta-data knowledge networks and members (in terms of tokens) of high-scoring topics is minimal on the string identity level, the entities that score highest with respect to node centrality metrics seemed to be suitable fits for topic labels. In future work, the appropriateness of this strategy for automatically finding labels for topics can be further explored. This strategy could supplement or replace the approach of using the most likely term per topic as the topic label, or identifying labels through a manual, interpretative process. As part of such extended work, it should also be tested how this effect correlates with term frequency.

Second, most of the key entities from the text-based knowledge networks also occur as topic members. This was observed for generic terms from the tested domains and datasets as well as for domain-specific terms. This relationship between topic members and key entities from text-based networks is asymmetric, i.e. outputs from topic modeling contain terms that do not occur among the key entities from the text-based networks. This is mainly due to the number of key entities retrieved (top ten) and their high overlap across network metrics such that the total set of key entities is smaller than the set of tokens across topics per text set. The analysis of the terms that rank high for topics but not among the key entities revealed that many of these terms were removed from the thesaurus (used for generating the networks on which key entity analysis was performed) by using the entity extractor built with supervised learning in chapter 3, as they were noisy or overly generic. This finding suggests that the key entities from networks constructed by using entity prediction models built with supervised learning (CRF) compare to the tokens identified via unsupervised learning (topic modeling) as follows when applied to the same inference data:

1. Similar terms are found via different term ranking methods, i.e. grouping words into sets of entities generated from the same topics (topic modeling) versus grouping nodes into sets of structurally entities (key entity analysis).
2. The same noise terms. This implies two more findings:
    - Topic modeling can benefit from the same cleaning techniques that were applied to the output of the entity extractor. Thus, the same delete lists and entity merger lists can be used for both outputs.
    - Applying the same cleaning techniques consistently to both output sets might further increase the similarity between the outcomes of both methods. The validity of this assumption needs to be tested in future work.

Third, even though the comparison between the key entities from the reference networks (meta-data and text-based) was not the focus of this study, a side-product of this chapter was finding out that for either network type, the key entities are highly similar across the considered network metrics. This finding further complements the outcome of the previous chapter by showing that key entities differ across network types, but are highly similar within networks constructed from the same data with either one method.

In summary, besides the proposition and testing of a methodological improvement to enhancing social network data with content nodes, a second contribution with this chapter was the comparison of the results from topic modeling; an efficient and unsupervised information extraction technique, to the outcome of alternative methods, including supervised entity extraction. Clearly, such comparisons cannot replace rigorous validations of topic modeling by comparing the results against ground truth data. However, such ground truth data might be expensive to collect. For example, for the Funding corpus, we might have some expertise in a few research domains, but are not qualified to evaluate topics from proposals from a wide range of areas over the last 18 years. Finding subject matter expert who are qualified to make these judgments would be expensive. Therefore, contrasting the outcome of topic modeling against alternative methods helps to understand the results of topic modeling in the wider context of information extraction methods. The comparisons in this chapter have led to the following conclusions: first, identifying content nodes from text-based knowledge networks by performing key player analysis retrieves only a small portion of entities that would not be found with topic modeling. Second, the key entities from meta-data knowledge networks might not only serve as fitting labels for topics, but might also be suitable proxies for some of the topics found with topic modeling. The validity of these assumptions needs to be tested in future work.

6.6 Limitations and Future Work

This chapter as well as the previous ones have shown that applying clearly defined information extraction methods still involve a plethora of decisions to make; many of which impact network analysis results. In this chapter, nodes had to be grouped into partitions. However, grouping is a science and art of its own, and is not the focus of this chapter. No tests were conducted if the retrieved groups are meaningful. This kind of assessment would require comparing the retrieved groups against ground truth data on meaningful partitions, which only network participants or subject matter experts could provide. Also, the grouping algorithm used herein as well the other common grouping techniques are defined for symmetric data. Since the network data used in this study are not symmetric, they had to be symmetrized prior to grouping. The same limitation, i.e. adjusting the actual characteristics of the data to the properties required for a computational routine or metric, also applies to most of the network metrics used in this thesis; with many of them being defined for squared, undirected, and binary matrices (see Table 154 for this information). Most software tools automatically internally transform network data such that they are compatible with the requirement for a routine, including ORA.

Another limitation that has also been observed in a prior chapter (4) is the incompatibility of tools: the original Funding data are represented in UTF8 encoding. Therefore, I used the same encoding for the relational database in which I managed the Funding data. However, ORA uses ASCII encoding, and converted non-ASCII letters into other symbols. Importing networks into ORA caused changes in the spelling of some agent nodes, and these altered names might not match database entries anymore. This is problematic for retrieving texts per person in order to put together the text sets per group. Such incompatibilities are not always obvious. Generating a mapping between encodings, which has not been done for this project, might have yet more impact on the analysis results.

With respect to the main point this chapter, i.e. improving the enhancement of social network data with content nodes; the following methodological extensions might be relevant for future work:

First, the identification of content nodes per group was done on a case-by-case basis for the largest groups per dataset and time period. This process can be standardized by automatically performing the following steps: pick the top N nodes from the top N topics, assign a generic or specific label per topic, e.g. the high-scoring term or a key entity from meta-data networks, and fuse the knowledge network with the social network. While technically, this procedure can be added to ORA by re-using existing routines, the validity of this process needs to be further validated on more datasets and groups.

Second, the identification of content nodes can be performed not only on the level of positions, roles or groups of agents, but also on the corpus level. This extension could serve two purposes: first, comparing the outcome of grouping agent nodes by employing grouping algorithms from social network analysis against grouping agents based on shared content, i.e. topics that multiple people are involved in. McCallum et al. (2007) have shown that clustering agent nodes based on topic modeling can outperform clustering of agents based on partitioning social network data via grouping algorithms. However, in that work, only dyads between email senders and receivers were identified. This idea can be extended to larger groups. Second, the social network could be enhanced with links between agents who are associated with the same content, but have not yet co-authored a document. This step serves three purposes: verifying existing links between agents, identifying missing links between agents, and suggesting additional ties between agent nodes as well as knowledge nodes. This last step would also allow for adding the impact of language use on network data into the network. However, further studies are needed to test for the validity of this approach.

Third, based on the conclusions from this chapter, it seems worthwhile to test the appropriateness of using key entities from meta-data networks as labels for topics in a more rigorous fashion and on additional datasets. This type of comparison can also serve another purpose: when topic modeling is performed on a per document basis, the identified topics can be manually labeled, and the resulting labels then compared against keywords the authors had selected per document. This comparison could help to understand the agreement or mismatch between top-down categorizations of documents, e.g. via pre-defined or self-defined keywords, versus bottom-up classifications of documents that emerge from the content of the text data.

7 Summary

This thesis makes novel contributions to the understanding of the impact of methodological choices that need to be made when coding texts as networks on the resulting network data. I show how network analysis results differ depending on the selection of (parameters for) a selected set of commonly used manual, computer-assisted and fully automated methods for information and relation extraction. The gained knowledge about the robustness of these methods and the propagation of selected types of errors can help researchers and practitioners to draw valid and reasonable conclusions from their analysis results and work presented by others.

The findings from analyzing the impact of a) reference resolution and b) proximity-based link formation on network data (chapter 2) emphasize the importance and practical relevance of understanding the amount and nature of the impact of both routines. Combining these insights with the outcomes of testing the prediction quality of an entity extractor (built and evaluated in chapter 3) in different applications settings (chapter 5) complements traditional methods for assessing the accuracy of several common relation extraction methods. Bringing the constructed prediction models into application contexts for which no ground truth data might be available (chapter 5) is particularly relevant as this resembles typical, real-world analysis scenarios. However, certain challenges arise when transitioning from experimental results and prediction models evaluated with classic assessment techniques to practical applications. These problems and their implications are identified, and solutions for mitigating them are developed, implemented and tested (chapter 4). A synthesis of lessons learned from this process is provided herein (end of chapter 4). Based on the gained understanding and these solutions, the results for using various methods for the end-to-end process of going from texts to networks are identified and compared (chapters 5, 6). This comparison leads to the suggestion of strategies for combining a selection of these methods such that a more comprehensive understanding of the underlying network can be gained (chapters 5, 6). For each project, applicable limitations and directions for future work are outlined at the end of each chapter.

Analyzing strategies for the joint consideration of text data and network data via enhancing network data with nodes that represents salient information from text data are developed by drawing from social network theory. The developed solution is tested on the same datasets as used for aforementioned applications scenarios. The presented work on linking groups or clusters of people to information that these people produce, share or process shows that extracting content nodes from documents from groups of *structurally equivalent* agents is an appropriate strategy for enabling the *comparison* of information, while extracting content nodes from groups of *structurally coherent* agents is appropriate for enabling the *enhancement* of social network data (chapter 6). The latter approach is not only brought into application contexts for evaluation, but aligns with the design of the studies from the previous chapters in that the outcomes of using topic modeling for identifying content nodes is compared to the results from the previously tested methods for relation extraction.

The findings from the presented experiments, evaluation studies and qualitative, in-depth analyses as well as the technology provided with this thesis can help people to collect, manage and analyze rich sets of network data at any scale. The availability of such data is a precondition for testing hypotheses, answering questions and advancing theories about networks. Overall, in this thesis, an interdisciplinary and computationally rigorous approach is used for bringing text data and relational data closer together; thereby advancing the intersection of network analysis, natural language processing and computer science.

References

Adamic, L., & Huberman, B. (1999). Growth dynamics of the world wide web. *Nature, 401*(6749), 131.

Adar, E., & Adamic, L. (2005, September). *Tracking Information Epidemics in Blogspace.* Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Compiegne, France, pp. 207-214.

Alderson, D. (2008). Catching the'network science'bug: Insight and opportunity for the operations researcher. *Operations Research, 56*(5), 1047-1065.

Allen, J., & Frisch, A. (1982). *What's in a semantic network?* Proceedings of 20th annual meeting of Association for Computational Linguistics (ACL), Toronto, ON, Canada, pp. 19-27.

Anderson, B. S., Butts, C., & Carley, K. M. (1999). The interaction of size and density with graph-level indices. *Social Networks, 21*(3), 239-268.

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen, 59*, 74-76.

Aven, B. L. (2010). *The Network Structure of Corrupt Innovation: The Case of Enron.* (PhD Thesis), Stanford University.

Bainbridge, W. S. (2007). The scientific research potential of virtual worlds. *Science, 317*(5837), 472-476.

Baker, W., & Faulkner, R. (1993). The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. *American Sociological Review, 58*(6), 837-860.

Balasubramanyan, R., Lin, F., & Cohen, W. W. (2010, December). *Node Clustering in Graphs: An Empirical Study.* Paper presented at Neural Information Processing Systems (NIPS) Workshop Networks Across Disciplines, Vancouver, Canada, pp. 1-8.

Barabási, A., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science, 286*(5439), 509-512.

Barthelemy, M., Chow, E., & Eliassi-Rad, T. (2005, March). *Knowledge Representation. Issues in Semantic Graphs for Relationship Detection.* Proceedings of AAAI Spring Symposium on AI Technologies for Homeland Security, Stanford, CA, pp. 91-98.

Bearman, P., & Stovel, K. (2000). Becoming a Nazi: A model for narrative networks. *Poetics, 27*(2-3), 69-90.

Bengtson, E., & Roth, D. (2008, October). *Understanding the value of features for coreference resolution.* Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Honolulu, HI, pp. 294-303.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., McCarty, C., Shelley, G. A., & Robinson, S. (1990). Comparing four different methods for measuring personal social networks. *Social Networks, 12*(3), 179-215.

Bernard, H. R., & Ryan, G. W. (1998). Text analysis: Qualitative and quantitative methods. In H. R. Bernard (Ed.), *Handbook of methods in cultural anthropology* (pp. 595–646). Walnut Creek: Altamire.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5), 34-43.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 1-36.

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997, March/ April). *Nymble: a high-performance learning name-finder.* Proceedings of Fifth conference on Applied Natural Language Processing (ANLP), Washington, DC, pp. 194-201.

Bikel, D. M., Schwartz, R., & Weischedel, R. (1999). An Algorithm that Learns What‘s in a Name. *Machine Learning, 34*(1-3), 211-231.

Biocca, Z., & Biocca, F. (2002). Building bridges across fields, universities, and countries: Successfully funding communication research through interdisciplinary collaboration. *Journal of Applied Communication Research, 30*(4), 350-357.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology, 92*(5), 1170-1182.

Bond, D., Bond, J., Oh, C., Jenkins, J., & Taylor, C. (2003). Integrated data for events analysis (IDEA): an event typology for automated events data development. *Journal of Peace Research, 40*(6), 733-745.

Borgatti, S. P. (2003). The key player problem. In R. Breiger, K. M. Carley & P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (pp. 241-252): National Academy of Sciences Press.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks, 28*(2), 124-136.

Borthwick, A. (1999). *A maximum entropy approach to named entity recognition.* (PhD Thesis), New York University.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998, August). *Exploiting diverse knowledge sources via maximum entropy in named entity recognition.* Proceedings of Sixth Workshop on Very Large Corpora, Montreal, Quebec, Canada, pp. 152-160.

Boster, J. S., Johnson, J. C., & Weller, S. C. (1987). Social position and shared knowledge: Actors' perceptions of status, role, and social structure. *Social Networks, 9*(4), 375-387.

Bourdieu, P. (1991). *Language and symbolic power*. Cambridge, MA: Harvard University Press.

Brandes, U., & Erlebach, T. (2005). *Network Analysis: Methodological Foundations*. Berlin, Heidelberg: Springer.

Breiger, R. L., Boorman, S. A., & Arabie, P. (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology, 12*(3), 328-383.

Brin, S. (1999, March). *Extracting Patterns and Relations from the World Wide Web.* Proceedings of The World Wide Web and Databases (WebDB) Workshop at EDBT, Valencia, Spain, pp. 172 - 183.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), 467-479.

Bunescu, R. (2007). *Learning for Information Extraction from Named Entity Recognition and Disambiguation to Relation Extraction.* (PhD Thesis), University of Texas.

Bunescu, R., Ge, R., Kate, R., Marcotte, E., Mooney, R., Ramani, A., & Wong, Y. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine, 33*(2), 139-155.

Bunescu, R., & Mooney, R. (2007). Statistical Relational Learning for Natural Language Information Extraction. In L. Getoor & B. Taskar (Eds.), *Statistical Relational Learning* (pp. 535 - 552). Cambridge, MA: MIT Press.

Bureau_of_Intelligence_and_Research. (2011). *Independent States in the World. Fact Sheet.* Washington, DC: Retrieved from http://www.state.gov/s/inr/rls/4250.htm.

Burt, R. S. (1976). Positions in networks. *Social Forces, 55*(1), 93-122.

Burt, R. S. (1999). The social capital of opinion leaders. *The Annals of the American Academy of Political and Social Science, 566*(1), 37-54.

Burt, R. S., & Janicik, G. A. (1996). Social contagion and social structure. In D. Iacobucci (Ed.), *Networks in marketing* (pp. 32-49). Thousand Oaks, CA: Sage.

Burt, R. S., & Lin, N. (1977). Network Time Series from Archival Records. In D. R. Heise (Ed.), *Sociological Methodology* (Vol. 1977, pp. 224-254). San Francisco, CA: Jossey-Bass.

Buzan, T. (1974). *Using both sides of the brain*. New York: Dutton.

Cafarella, M., Banko, M., & Etzioni, O. (2006). *Relational web search*. Technical Report, 06-04-02, University of Washington.

Carley, K. M. (1988). Formalizing the Social Expert's Knowledge. *Sociological Methods & Research, 17*(2), 165-232.

Carley, K. M. (1991). Designing organizational structures to cope with communication breakdowns: a simulation model. *Industrial Crises Quarterly, 5*(1), 19-57.

Carley, K. M. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology, 23*, 75-126.

Carley, K. M. (1994). Extracting culture through textual analysis. *Poetics, 22*, 291-312.

Carley, K. M. (1997a). Extracting team mental models through textual analysis. *Journal of Organizational Behavior, 18*, 533-558.

Carley, K. M. (1997b). Network text analysis: The network position of concepts. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts* (pp. 79–100). Mahwah, NJ: Lawrence Erlbaum Associates.

Carley, K. M. (2002a). Smart agents and organizations of the future. In L. Lievrouw & S. Livingstone (Eds.), *The Handbook of New Media: Social Shaping and Consequences of ICTs* (pp. 206–220). Thousand Oaks, CA: Sage.

Carley, K. M. (2002b). Summary of key network measures for characterizing organizational architectures. *Unpublished Document: CMU.*

Carley, K. M. (n.a.). A Structural Approach to the Incorporation of Cultural Knowledge in Adaptive Adversary Models - Overview Retrieved 05/17/2011, from http://www.casos.cs.cmu.edu/projects/project.php?ID=18&Name=A%20Structural%20Approach%20to%20the%20Incorporation%20of%20Cultural%20Knowledge%20in%20Adaptive%20Adversary%20Models

Carley, K. M., Columbus, D., Bigrigg, M., & Kunkel, F. (2011). *AutoMap User's Guide 2011*. Technical Report, CMU-ISR-11-108, Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Carley, K. M., Diesner, J., Reminga, J., & Tsvetovat, M. (2007). Toward an interoperable dynamic network analysis toolkit. *Decision Support Systems. Special Issue Cyberinfrastructure for Homeland Security, 43*(4), 1324-1347.

Carley, K. M., & Kaufer, D. S. (1993). Semantic Connectivity: An Approach for Analyzing Symbols in Semantic Networks. *Communication Theory, 3*(3), 183-213.

Carley, K. M., Lanham, M., Martin, M., Morgon, G., Schmerl, B., van Holt, T., . . . Kunkel, F. (2011, February). *Rapid Ethnographic Assessment: Data to Model.* Paper presented at HSCB Focus 2011: Integrating Social Science Theory and Analytic Methods for Operational Use, Chantilly, VA.

Carley, K. M., Lee, J. S., & Krackhardt, D. (2001). Destabilizing networks. *Connections, 24*(3), 31-34.

Carley, K. M., & Palmquist, M. (1991). Extracting, Representing, and Analyzing Mental Models. *Social Forces, 70*(3), 601 - 636.

Carley, K. M., Reminga, J., Storrick, J., & Columbus, D. (2011). *ORA User's Guide 2011*. Technical Report, CMU-ISR-11-107, Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and Methods in Social Network Analysis*. Cambridge, New York: Cambridge University Press.

Cataldo, M., & Herbsleb, J. D. (2008, November). *Communication networks in geographically distributed software development.* Proceedings of ACM conference on Computer Supported Cooperative Work (CSCW), San Diego, CA, pp. 579-588.

Cataldo, M., Herbsleb, J. D., & Carley, K. M. (2008). *Socio-technical congruence: a framework for assessing the impact of technical and work dependencies on software development* *productivity.* Proceedings of 2nd ACM-IEEE international symposium on empirical software engineering and measurement (ESEM), Kaiserslautern, Germany, pp. 2-11.

Cataldo, M., Wagstrom, P. A., Herbsleb, J. D., & Carley, K. M. (2006, November). *Identification of coordination requirements: implications for the Design of collaboration and awareness tools.* Proceedings of 20th Conference on Computer Supported Cooperative Work (CSCW), Banff, Canada, pp. 353-362.

Central_Intelligence_Agency. (2009). The World Factbook, 2011, from https://www.cia.gov/library/publications/the-world-factbook/

Chang, J., Boyd-Graber, J., & Blei, D. (2009, June/ July). *Connections between the lines: augmenting social networks with text.* Proceedings of 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Paris, France, pp. 169-178.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009, December). *Reading Tea Leaves: How Humans Interpret Topic Models.* Proceedings of Neural Information Processing Systems (NIPS), Vancouver, Canada.

Chinchor, N. (2001). *Message Understanding Conference (MUC) 7.* Linguistic Data Consortium, Philadelphia.

Chinchor, N., & Sundheim, B. (2003). *Message Understanding Conference (MUC) 6*. Linguistic Data Consortium, Philadelphia.

Ciaramita, M., & Altun, Y. (2005, December). *Named-entity recognition in novel domains with external lexical knowledge.* Paper presented at Neural Information Processing Systems (NIPS) Workshop on Advances in Structured Learning for Text and Speech Processing, Whistler, BC, Canada.

Cohen, W. W., Ravikumar, P., & Fienberg, S. (2003, August). *A comparison of string metrics for matching names and records.* Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD) Workshop on Data Cleaning and Object Consolidation, Washington, DC, pp. 73-78.

Cohen, W. W., & Sarawagi, S. (2004, August). *Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods.* Proceedings of 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Seattle, WA, pp. 89-98.

Coleman, J., Katz, E., & Menzel, H. (1966). *Medical innovation*. Indianapolis: Bobbs-Merrill.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407-428.

Collins, A., & Quillian, M. (1969). Retrieval Time from Semantic Memory. *Journal of Verbal Learning & Verbal Behavior, 8*(2), 240-248.

CoNLL-2003. (2003). Language-Independent Named Entity Recognition (II), from http://www.cnts.ua.ac.be/conll2003/ner/

CORDIS. Community Research and Development Information Service from http://cordis.europa.eu/search

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems: Centering Resonance Analysis of Communication. *Human Communication Research, 28*(2), 157-206.

Cowan, R., Jonard, N., & Zimmermann, J. (2003). The joint dynamics of networks and knowledge. In R. Cowan & N. Jonard (Eds.), *Heterogenous Agents, Interactions, and Economic Performance. Lecture Notes in Economics and Mathematical Systems* (Vol. 521, pp. 155-174): Springer.

Culotta, A., & McCallum, A. (2005). *Joint deduplication of multiple record types in relational data.* Proceedings of 14th ACM International Conference on Information and Knowledge Management (ICIKM), Bremen, Germany, pp. 257-258.

Cummings, J. N., & Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science, 35*(5), 703-722.

Dabbish, L., Towne, B., Diesner, J., & Herbsleb, J. D. (2011). Construction of association networks from communication in teams working on complex projects. *Statistical Analysis and Data Mining, 4*(5), 547-563.

Danowski, J. A. (1993). Network Analysis of Message Content. *Progress in Communication Sciences, 12*, 198-221.

Danowski, J. A., & Edison-Swift, P. (1985). Crisis effects on intraorganizational computer-based communication. *Communication Research, 12*(2), 251-270.

Daumé, H. (2007, June). *Frustratingly easy domain adaptation.* Proceedings of 45th Annual Meeting of the Association of Computational Linguistics (ACL), Prague, Czech Republic, pp. 256–263.

Deemter, K., & Kibble, R. (2000). On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics, 26*(4), 629-637.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science, 41*(6), 391-407.

Denis, P., & Baldridge, J. (2007, April). *Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming.* Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT), Rochester, NY, pp. 236-243.

Diesner, J., & Carley, K. M. (2004). *AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts*. Technical Report CMU-ISRI-04-100, Carnegie Mellon University, School of Computer Science, Institute for Software Research International.

Diesner, J., & Carley, K. M. (2005a, April). *Exploration of Communication Networks from the Enron Email Corpus.* Proceedings of SIAM International Conference on Data Mining, Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA, pp. 3-14.

Diesner, J., & Carley, K. M. (2005b). Revealing social structure from texts: Meta-Matrix text analysis as a novel method for network text analysis. In V. K. Narayanan & D. J. Armstrong (Eds.), *Causal Mapping for Information Systems and Technology Research:*

Approaches, Advances, and Illustrations (pp. 81-108). Harrisburg, PA: Idea Group Publishing.

Diesner, J., & Carley, K. M. (2008a). Conditional Random Fields for Entity Extraction and Ontological Text Coding. *Journal of Computational and Mathematical Organization Theory, 14*(3), 248 - 262.

Diesner, J., & Carley, K. M. (2008b). *Looking under the hood of machine learning algorithms for parts of speech tagging*. Technical Report CMU-ISR-08-131R, Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Diesner, J., & Carley, K. M. (2009a, July). *He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis.* Proceedings of IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA), Ottawa, Canada, pp. 9-16.

Diesner, J., & Carley, K. M. (2009b, April). WYSIWII - What You See Is What It Is: Informed Approximation of Relational Data from Texts. Presentation at *General Online Research (GOR) Conference*, Vienna, Austria.

Diesner, J., & Carley, K. M. (2010a, December). *Extension of supervised conditional machine learning system to extract socio-technical networks from open source text data about the Sudan*, Vancouver, Canada.

Diesner, J., & Carley, K. M. (2010b, August). *A methodology for integrating network theory and topic modeling and its application to innovation diffusion.* Proceedings of IEEE International Conference on Social Computing (SocialComp), Workshop on Finding Synergies Between Texts and Networks, Minneapolis, MN.

Diesner, J., & Carley, K. M. (2010c). Relation Extraction from Texts (in German, title: Extraktion relationaler Daten aus Texten). In C. Stegbauer & R. Häußling (Eds.), *Handbook Network Research (Handbuch Netzwerkforschung)* (pp. 507-521): Vs Verlag.

Diesner, J., & Carley, K. M. (2011a). Semantic Networks. In G. Barnett & J. G. Golson (Eds.), *Encyclopedia of Social Networking* (pp. 766-769). Thousand Oaks, CA: Sage.

Diesner, J., & Carley, K. M. (2011b). Words and Networks. In G. Barnett & J. G. Golson (Eds.), *Encyclopedia of Social Networking* (pp. 958-961). Thousand Oaks, CA: Sage.

Diesner, J., Carley, K. M., & Katzmair, H. (2007, May). *The morphology of a breakdown. How the semantics and mechanics of communication networks from an organization in crises relate.* Paper presented at XXVII Sunbelt Social Network Conference, Corfu, Greece.

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication Networks from the Enron Email Corpus. “It's Always About the People. Enron is no Different”. *Computational & Mathematical Organization Theory, 11*(3), 201-228.

Dietterich, T. G. (2002, August). *Machine Learning for Sequential Data: A Review.* Proceedings of Joint IAPR International Workshops SSPR 2002 and SPR 2002, Windsor, ON, Canada, pp. 15-33.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004, May). *The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation.*

Proceedings of 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, pp. 837–840.

Doerfel, M. (1998). What Constitutes Semantic Network Analysis? A Comparison of Research and Methodologies. *Connections, 21*(2), 16-26.

Doerfel, M., & Barnett, G. (1999). A Semantic Network Analysis of the International Communication Association. *Human Communication Research, 25*(4), 589-603.

Eagle, N., & Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing, 10*(4), 255-268.

Eckert, P. (2000). *Language variation as social practice*. Oxford: Blackwell Publishers.

Ehrlich, K., Lin, C., & Griffiths-Fisher, V. (2007, October). *Searching for experts in the enterprise: combining text and social network analysis.* Proceedings of 2007 ACM International Conference on Supporting Group Work (GROUP), Sanibel Island, FL, pp. 117-126.

Elageed, A. A. E. (2008). *Weaving the Social Networks of Women Migrants in Sudan: The Case of Gezira*: LIT Verlag.

Emery, F. E., & Trist, E. L. (1960). Socio-technical systems. In C. W. Churchman & M. Verhulst (Eds.), *Management Science Models and Techniques* (Vol. 2, pp. 83-97). London: Pergamon.

Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen, 6*, 290-297.

Erickson, B. (1981). Secret Societies and Social Structure. *Social Forces, 60*(1), 188-210.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., . . . Yates, A. (2004, May). *Web-scale information extraction in KnowItAll (Preliminary Results).* Proceedings of 13th International Conference on the World Wide Web (WWW), New York, NY, pp. 100-110.

Everett, M. G., & Borgatti, S. P. (1994). Regular equivalence: General theory. *Journal of mathematical sociology, 19*(1), 29-52.

Faust, K. (2006). Comparing social networks: size, density, and local structure. *Metodološki zvezki, 3*(2), 185-216.

Feldman, G. D., & Seibel, W. (2006). *Networks of Nazi persecution: bureaucracy, business, and the organization of the Holocaust*. New York: Berghahn Books.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fillmore, C. J. (1968). The Case for Case. In E. Bach & R. T. Harms (Eds.), *Universals in Linguistic Theory* (pp. 1-88). New York, NY: Holt, Rinehart, and Winston.

Fillmore, C. J. (1982). Frame Semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the morning calm* (pp. 111-137). Seoul, South Korea: Hanshin Publishing.

Fitzmaurice, S. (2000). Coalitions and the investigation of social influence in linguistic History. *European Journal of English Studies, 4*(3), 265-276.

Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003, May/ June). *Named entity recognition through classifier combination.* Proceedings of 7th Conference on Computational Natural Language Learning (CoNLL), Edmonton, Canada, pp. 168-171.

Folkstad, J., & Hayne, S. C. (2011, January). *Visualization and Analysis of Social Networks of Research Funding.* Proceedings of 44th Hawaii International Conference on System Sciences (HICSS), Kanaui, HI.

Fox, L. (2003). *Enron: The rise and fall*. Hoboken: Wiley.

Frantz, T. L., Cataldo, M., & Carley, K. M. (2009). Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational & Mathematical Organization Theory, 15*(4), 303-328.

Franzosi, R. (1989). From words to numbers: A generalized and linguistics-based coding procedure for collecting textual data. *Sociological Methodology, 19*(1990), 225-257.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215-239.

Freeman, L. C. (2004). *The development of social network analysis*. Vancouver, Canada: Empirical Press.

Friedkin, N. E. (1981). The development of structure in random networks: an analysis of the effects of increasing network density on five measures of structure. *Social Networks, 3*(1), 41-52.

Fusaro, P. C., & Miller, R. M. (2002). *What went wrong at Enron: Everyone's guide to the largest bankruptcy in US history*. Hoboken: Wiley.

Galaskiewicz, J., & Burt, R. S. (1991). Interorganization contagion in corporate philanthropy. *Administrative Science Quarterly, 36*(1), 88-105.

Gerdes, L. (2008). *Codebook for Network Data on Individuals Involved with Terrorism and Counterterrorism*. Techincal Report, CMU-ISR-08-136, Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Gerner, D. J., Schrodt, P. A., Francisco, R. A., & Weddle, J. L. (1994). Machine Coding of Event Data Using Regional and International Sources. *International Studies Quarterly, 38*(1), 91-119.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(11), 7821-7826.

Giuffre, K. (2001). Mental Maps: Social Networks and the Language of Critical Reviews. *Sociological Inquiry, 71*(3), 381-393.

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York, NY: Aldine.

Gloor, P., Krauss, J., Nann, S., Fischbach, K., Schoder, D., & Switzerland, B. (2009, August). *Web Science 2.0: Identifying Trends through Semantic Social Network Analysis.* Proceedings of IEEE Conference on Social Computing (SocialCom), Vancouver, Canada.

Gloor, P., & Zhao, Y. (2006, July). *Analyzing actors and their discussion topics by semantic social network analysis.* Proceedings of 10th IEEE International Conference on Information Visualisation London, UK.

Golder, S. A. (2003). *A typology of social roles in usenet.* (B.A. Thesis), Harvard University.

Goldstein, J. (1992). A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution, 36*(2), 369-385.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211-244.

Gumperz, J. (1982). Social Network and Language Shift. In J. Gumperz (Ed.), *Discourse Strategies* (pp. 38-58). Cambridge: Cambridge University Press.

Gupta, R., & Sarawagi, S. (2009). Domain adaptation of information extraction models. *ACM SIGMOD Record, 37*(4), 35-40.

Hachey, B., Grover, C., & Tobin, R. (2006). *Datasets for generic relation extraction. LDC2011T08*. Linguistic Data Consortium, Philadelphia.

Hämmerli, A., Gattiker, R., & Weyermann, R. (2006). Conflict and Cooperation in an Actors' Network of Chechnya Based on Event Data. *Journal of Conflict Resolution, 50*(2), 159-175.

Harper, W. R., & Harris, D. H. (1975). The application of link analysis to police intelligence. *Human Factors, 17*(2), 157-164.

Harrer, A., Malzahn, N., Zeini, S., & Hoppe, H. (2007). *Combining social network analysis with semantic relations to support the evolution of a scientific community.* Proceedings of 8th International Conference on Computer Supported Collaborative Learning (CSCL), New Brunswick, NJ, pp. 270-279.

Hartley, R., & Barnden, J. (1997). Semantic networks: visualizations of knowledge. *Trends in Cognitive Sciences, 1*(5), 169-175.

Haythornthwaite, C. (2001). Exploring multiplexity: Social network structures in a computer-supported distance learning class. *The Information Society, 17*(3), 211-226.

Haythornthwaite, C., & Gruzd, A. (2008, May). *Analyzing networked learning texts.* Proceedings of 6th International Conference on Networked Learning, Halkidiki, Greece, pp. 136-143.

Hendrickx, I., Kim, S., Kozareva, Z., & Nakov, P. (2009, June). *Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals.* Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) Workshop on Semantic Evaluations: Recent Achievements and Future Directions, Boulder, CO, pp. 94-99.

Hirst, G. (2006). Views of text-meaning in computational linguistics: Past, present, and future. In G. Dodig-Crnkovic & S. Stuart (Eds.), *Computation, Information, Cognition–The Nexus and the Liminal* (pp. 270–279). Cambridge: Cambridge Scholars Publishing.

Hobbs, J. (1979). Coherence and coreference. *Cognitive science, 3*(1), 67-90.

Horta, H., Huisman, J., & Heitor, M. (2008). Does competitive research funding encourage diversity in higher education? *Science and Public Policy, 35*(3), 146-158.

Hovy, E. H. (1990). *Parsimonious and profligate approaches to the question of discourse structure relations.* Proceedings of 5th International Workshop on Natural Language Generation, Dawson, PA, pp. 128-136.

Howard, R. (1989). Knowledge maps. *Management Science, 35*(8), 903-922.

Howlett, J. (1980). Analytical Investigative Techniques: Tools for Complex Criminal Investigations. *Police Chief, 47*(12), 42-45.

Hummon, N., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*(1), 39-63.

Humphreys, M. (2005). Natural resources, conflict, and conflict resolution: Uncovering the mechanisms. *Journal of Conflict Resolution, 49*(4), 508-537.

Janas, J., & Schwind, C. (1979). Extensional Semantic Networks. In N. V. Findler (Ed.), *Associative Networks. Representation and Use of Knowledge by Computers.* (pp. 267 - 302). New York, San Francisco, London: Academic Press.

Johnson, J. C., Boster, J. S., & Palinkas, L. A. (2003). Social roles and the evolution of networks in extreme and isolated environments. *Journal of Mathematical Sociology, 27*(2-3), 89-121.

Johnson, J. C., & Krempel, L. (2004). Network Visualization: The" Bush Team" in Reuters News Ticker 9/11-11/15/01. *Journal of social structure, 5*.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.): Prentice Hall.

Kamp, H. (1981). A Theory of Truth and Semantic Representation Formal Methods in the Study of Language. In J. A. G. Groenendijk, T. M. V. Janssen & M. B. J. Stokhof (Eds.), *Formal Methods in the Study of Language.* (pp. 277-322): Mathematical Centre Tracts 135, Amsterdam.

Katz, E., & Lazarsfeld, P. F. (1955). *Personal influence*. New York: Free Press.

Kauffman, S. (1995). *At home in the universe: The search for laws of self-organization and complexity*. New York: Oxford University Press.

Keegan, B., Ahmed, M. A., Williams, D., Srivastava, J., & Contractor, N. (2010, August). *Dark Gold: Statistical Properties of Clandestine Networks in Massively Multiplayer Online Games.* Proceedings of Social Computing (SocialCom), Minneapolis, MN, pp. 201-208.

King, G., & Lowe, W. (2003). An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization, 57*(3), 617-642.

Klein, D., Smarr, J., Nguyen, H., & Manning, C. (2003, May/ June). *Named Entity Recognition with Character-Level Models.* Proceedings of Conference on Computational Natural Language Learning (CoNLL), Edmonton, Canada.

Kleinberg, J. (2003). Bursty and Hierarchical Structure in Streams. *Data Mining and Knowledge Discovery, 7*(4), 373-397.

Klerks, P. (2001). The network paradigm applied to criminal organizations: theoretical nitpicking or a relevant doctrine for investigators. *Connections, 24*(3), 53-65.

Klimoski, R., & Mohammed, S. (1994). Team Mental Model: Construct or Metaphor? *Journal of Management, 20*(2), 403.

Knoke, D., & Yang, S. (2008). *Social network analysis*. Thousand Oaks, CA: Sage.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks, 9*, 109-134.

Krackhardt, D. (1990). Assessing the political landscape: Structure, cognition, and power in organizations. *Administrative Science Quarterly, 35*, 342-369.

Krackhardt, D. (1994). Graph theoretical dimensions of informal organizations. *Computational organization theory, 89*(112), 123-140.

Krackhardt, D. (1998). Simmelian ties: Super strong and sticky. In R. Kramer & M. Neale (Eds.), *Power and influence in organizations* (pp. 21-38). Thousand Oaks, CA: Sage.

Krackhardt, D., & Carley, K. M. (1998, June). *A PCANS Model of Structure in Organization.* Paper presented at International Symposium on Command and Control Research and Technology, Monterray, CA, pp. 113-119.

Kraut, R., Rice, R., Cool, C., & Fish, R. (1998). Varieties of social influence: The role of utility and norms in the success of a new communication medium. *Organization Science, 9*(4), 437-453.

Krebs, V. (2002). Mapping networks of terrorist cells. *Connections, 24*(3), 43-52.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Kruskal, J. B. (1977). The relationship between multidimensional scaling and clustering. In J. V. Ryzin (Ed.), *Classification and clustering* (pp. 17-44). New York: Academic Press.

Lafferty, J., McCallum, A., & Pereira, F. (2001, June). *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* Proceedings of 18th International Conference on Machine Learning, Williamstown, MA, pp. 282-289.

Lampe, C. A. C., Ellison, N., & Steinfield, C. (2007, April/ May). *A familiar face (book): profile elements as signals in an online social network.* Proceedings of Computer Human Interaction Conference (CHI), San Jose, CA, pp. 435-444.

Landwehr, P., Diesner, J., & Carley, K. M. (2009, September). *Words of Warcraft: a relational text analysis of quests in an MMORPG.* Proceedings of Digital Games Research Association Conference (DiGRA), London, UK.

Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters, 87*(19), 198701.

Leinhardt, S. (1977). *Social networks: A developing paradigm*. New York: Academic Press.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009, June/ July). *Meme-tracking and the dynamics of the news cycle.* Proceedings of 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Paris, France, pp. 497–506.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*(1).

Leung, R. (2007). Network position, research funding and interdisciplinary collaboration among nanotechnology scientists: An application of social network analysis. *Solid State Phenomena, 121*, 1347-1350.

Lewins, A., & Silver, C. (2007). *Using Software in Qualitative Research: A Step-by-step Guide*. London, UK: Sage.

Lewis, M. P. (2009). Ethnologue: Languages of the World Sixteenth Edition, from http://www.ethnologue.com

Linguistic_Data_Consortium. (2005). *ACE English Annotation Guidelines for Relations*.

Lippi-Green, R. (1989). Social network integration and language change in progress in a rural alpine village. *Language in society, 18*(2), 213-234.

Lobban, R. (1975). Alienation, Urbanisation, and Social Networks in the Sudan. *The Journal of Modern African Studies, 13*(03), 491-500.

Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *Journal of mathematical sociology, 1*(1), 49-80.

Malm, A., Kinney, J., & Pollard, N. (2008). Social Network and Distance Correlates of Criminal Associates Involved in Illicit Drug Production. *Security Journal, 21*(1-2), 77-94.

Mandel, M. J. (1983). Local roles and social networks. *American sociological review, 48*, 376-386.

Marcoccia, M. (2004). On-line polylogues: conversation structure and participation framework in internet newsgroups. *Journal of Pragmatics, 36*(1), 115-145.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology, 16*, 435-463.

Mayfield, J., McNamee, P., & Piatko, C. (2003, May/ June). *Named entity recognition using hundreds of thousands of features.* Proceedings of Seventh conference of Natural Language Learning at Human Language Technologies and North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, pp. 184-187.

McAllister, I., & Studlar, D. (1991). Bandwagon, underdog, or projection?: Opinion polls and electoral choice in Britain, 1979-1987. *Journal of Politics, 53*(3), 720 - 740.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit, from http://mallet.cs.umass.edu

McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. *ACM Queue, 3*(9), 48-57.

McCallum, A., & Li, W. (2003, May/ June). *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.* Proceedings of Seventh conference of Natural Language Learning at Human Language Technologies and North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, pp. 188-191.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research, 30*, 249-272.

McCallum, A., Wang, X., & Mohanty, N. (2007). Joint Group and Topic Discovery from Relations and Text *Statistical Network Analysis: Models, Issues, and New Directions. Lecture Notes in Computer Science 4503* (pp. 28-44): Springer.

McClelland, C. A. (1971). *The Management and Analysis of International Event Data: A Computerized System for Monitoring and Projecting Event Flows*. AD0731074, University of Southern California, School of International Relations.

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*(1), 415-444.

Melkers, J., & Wu, Y. (2009). Evaluating the Improved Research Capacity of EPSCoR States: R&D Funding and Collaborative Networks in the NSF EPSCoR Program. *Review of Policy Research, 26*(6), 761-782.

Merrill, J., Bakken, S., Rockoff, M., Gebbie, K., & Carley, K. M. (2007). Description of a method to support public health information management: organizational network analysis. *Journal of Biomedical Informatics, 40*(4), 422-428.

Merton, R. K. (1968). *Social theory and social structure*. New York: Free Press.

Mihalcea, R. F., & Radev, D. R. (2011). *Graph-based Natural Language Processing and Information Retrieval*. Cambridge, UK: Cambridge University Press.

Milgram, S. (1967). The small world problem. *Psychology today, 1*(1), 60-67.

Miller, S., Fox, H., Ramshaw, L., & Weischedel, R. (2000, April/ May). *A novel use of statistical parsing to extract information from text.* Proceedings of 1st Conference of North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, WA, pp. 226-233.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics, 21*(2), 339-384.

Milroy, L. (1987). *Language and social networks* (2nd ed.). Oxford: Blackwell.

Mimno, D., & McCallum, A. (2008, July). *Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression.* Proceedings of 24th Conference on Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland, pp. 411-418.

Minsky, M. (1974). A Framework for Representing Knowledge. Presentation at MIT-AI Laboratory Memo 306.

Mitchell, A., Strassel, S., Huang, S., & Zakhary, R. (2005). *ACE 2004 Multilingual Training Corpus. LDC2005T09*. Linguistic Data Consortium, Philadelphia

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., . . . Sundheim, B. (2004). *TIDES Extraction (ACE) 2003 Multilingual Training Data. LDC2004T09*. Linguistic Data Consortium, Philadelphia.

Mitchell, J. (1969). The concept and use of social networks. In J. Mitchell (Ed.), *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns* (pp. 1-50). University of Manchester: University Press.

Mitchell, P. M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313-330.

Mohr, J. (1998). Measuring Meaning Structures. *Annual Reviews in Sociology, 24*(1), 345-370.

Monge, P. R., & Contractor, N. (2003). *Theories of Communication Networks*. New York: Oxford University Press.

MUC7. (2001). Named Entity Scores - English, from http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_english_score_report.html

Nadel, S. F. (1957). *The theory of social structure*. Glencoe: Free Press.

Nastase, V., & Szpakowicz, S. (2003, January). *Exploring noun-modifier semantic relations.* Proceedings of Fifth International Workshop on Computational Semantics (IWCS), Tilburg, Netherlands, pp. 285–301.

National_Research_Council. (2005). *Network Science*: National Academies Press.

Newcomb, T. (1961). *The acquaintance process*. New York: Holt, Rinehart and Winston.

Newman, M. E. J. (2010). *Networks: an introduction*. Oxford: Oxford University Press.

Newman, M. E. J., Barabasi, A. L., & Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton, NJ: Princeton University Press.

Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E, 64*(2), 26118.

Novak, J., & Cañas, A. (2008). *The Theory Underlying Concept Maps and How to Construct Them*. Florida Institute for Human and Machine Cognition, IHMC CmapTools Rev 01-2008, IHMC CmapTools 2006-01 Rev 01-2008,

Novak, J., & Gowin, D. (1984). *Learning How to Learn*. New York: Cambridge University Press.

NSF. National Science Foundation, from http://www.nsf.gov/

Ogden, C., & Richards, I. (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism* London: Routledge & Kegan Paul.

Palmquist, M., Carley, K. M., & Dale, T. A. (1997). Applications of Computer-Aided Text Analysis: Analyzing Literary and Nonliterary Texts. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences* (pp. 171-190). Mahwah, NJ: Lawrence Erlbaum Associates.

Paolillo, J. (1999). The virtual speech community: Social network and language variation on IRC. *Journal of Computer-Mediated Communication, 4*(4).

Parastatidis, S., Viegas, E., & Hey, T. (2009). Viewpoint: Smart Cyberinfrastructure for Research. A view of semantic computing and its role in research. *Communications of the ACM, 52*(12), 33-37.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pereira, D. V. (2004). *Automatic Lexicon Generation for Unsupervised Part-of-Speech Tagging Using Only Unannotated Text.* (Master Thesis), Virginia Polytechnic Institute and State University, Falls Church, VA.

Pfeffer, J., & Carley, K. M. (under review). Rapid Modeling and Analyzing Networks Extracted from Pre-Structured News Articles. Presentation at

Popping, R. (2003). Knowledge Graphs and Network Text Analysis. *Social Science Information, 42*(1), 91-106.

Powers, W. C., Troubh, R. S., & Winokur, H. S. (2002). *Report of investigation by the special investigative committee of the board of directors of Enron Corp*.

Qureshi, P. A. R., Memon, N., Wiil, U. K., & Karampelas, P. Detecting Social Polarization and Radicalization. *International Journal of Machine Learning and Computing, 1*(1), 49-57.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C. (2010, October). *A multi-pass sieve for coreference resolution.* Proceedings of Empirical Methods in Natural Language Processing (EMNLP), Boston, MA, pp. 492-501.

Ratinov, L., & Roth, D. (2009, June). *Design challenges and misconceptions in named entity recognition.* Proceedings of 13th Conference on Computational Natural Language Learning (CoNLL), Boulder, CO, pp. 147-155.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., . . . Versley, Y. *SemEval-2010 Task 1: Coreference resolution in multiple languages. LDC2011T01*. Linguistic Data Consortium, Philadelphia.

Reiss, A. J. (1988). Co-offending and Criminal Careers. In N. Morris & M. Tonry (Eds.), *Crime and Justice* (Vol. 10, pp. 117-170). Chicago: Chicago University Press.

Richards, T. (2002). An intellectual history of NUD* IST and NVivo. *International Journal of Social Research Methodology, 5*(3), 199-214.

Richards, W. D. (1971). *An Improved Conceptually-Based Method for Analysis of Communication Network Structure of Large Complex Organizations*. Michigan State University, Dept. of Communication.

Richards, W. D., & Rice, R. E. (1981). The NEGOPY network analysis program. *Social Networks, 3*(3), 215-223.

Roberts, C. W. (1997a). A Generic Semantic Grammar for Quantitative Text Analysis: Applications to East and West Berlin Radio News Content from 1979. *Sociological Methodology, 27*(1), 89-129.

Roberts, C. W. (1997b). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Roberts, K. H., & O'Reilly III, C. A. (1979). Some correlations of communication roles in organizations. *Academy of Management Journal, 22*(1), 42-57.

Rogers, E. (1962). *Diffusion of Innovations*. Glencoe: Free Press.

Roth, C. (2006, September). *Binding social and semantic networks.* Paper presented at 2nd European Conference on Complex Systems (ECCS) Oxford, UK.

Roth, C., & Cointet, J. (2010). Social and semantic coevolution in knowledge networks. *Social Networks, 32*(1), 16-29.

Roth, D., & Yih, W. (2002, August/ September). *Probabilistic reasoning for entity and relation recognition.* Proceedings of International Conference on Computational Linguistics (COLING), Taipei, Taiwan.

Roth, D., & Yih, W. (2007). Global Inference for Entities and Relations Identification via a Linear Programming Formulation. In L. Getoor & B. Taskar (Eds.), *Statistical Relational Learning* (pp. 535-552). Boston, MA: MIT Press.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box; prospects and limits in the search for mental models. *Psychological Bulletin*(100), 349-363.

Ryan, B., & Gross, N. (1943). The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology, 8*(1), 15-24.

Ryan, G. W., & Bernard, H. R. (2000). Data management and analysis methods. *Handbook of qualitative research, 2*, 769-802.

Saaty, T. L. (2005). *Theory and Applications of analytic network process* (Vol. 4922): RWS publications Pittsburgh, PA.

Sageman, M. (2004). *Understanding Terror Networks*. Philadelphia, PA: University of Pennsylvania Press.

Sailer, L. D. (1979). Structural equivalence: Meaning and definition, computation and application. *Social Networks, 1*(1), 73-90.

Sampson, S. (1968). *A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships.* (PhD Thesis), Cornell University.

Sandhaus, E. (2008). *The New York Times Annotated Corpus. LDC2008T19*. Linguistic Data Consortium, Philadelphia.

Sarawagi, S.). CRF Project Page (http://crf.sourceforge.net/). Presentation at

Sarawagi, S. (2008). Information extraction. *Foundations and Trends in Databases, 1*(3), 261-377.

Sarnecki, J. (2001). *Delinquent Networks: Youth Co-Offending in Stockholm*. Cambridge, UK: Cambridge University Press.

Satpal, S., & Sarawagi, S. (2007, September). *Domain Adaptation of Conditional Probability Models Via Feature Subsetting.* Proceedings of 11th European conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007), Warsaw, Poland, pp. 224-235.

Schrodt, P. A. (2001, February). *Automated coding of international event data using sparse parsing techniques.* Paper presented at International Studies Association, Chicago, Il.

Schrodt, P. A., Gerner, D. J., & Yilmaz, Ö. (2004, March). *Using Event Data to Monitor Contemporary Conflict in the Israel-Palestine Dyad.* Paper presented at International Studies Association, Montreal, Quebec, Canada.

Schrodt, P. A., Simpson, E. M., & Gerner, D. J. (2001, June). *Monitoring conflict using automated coding of newswire reports: a comparison of five geographical regions.* Paper presented at PRIO/Uppsala University/DECRG High-Level Scientific Conference on Identifying Wars: Systematic Conflict Research and Its Utility in Conflict Resolution and Prevention, Uppsala, Sweden.

Schrodt, P. A., Yilmaz, Ö., Gerner, D. J., & Hermick, D. (2008, March). *Coding Sub-State Actors using the CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework.* Paper presented at Annual Meeting of the International Studies Association, San Francisco, CA.

Seibel, W., & Raab, J. (2003). Verfolgungsnetzwerke. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie, 55*(2), 197-230.

Sha, F., & Pereira, F. (2003, May/ June). *Shallow Parsing with Conditional Random Fields.* Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL), Edmonton, Canada.

Shahaf, D., & Guestrin, C. (2010). *Connecting the dots between news articles.* Proceedings of 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 623-632.

Shanahan, J. G., Qu, Y., & Wiebe, J. M. (2006). *Computing attitude and affect in text: theory and applications*. New York: Springer.

Shapiro, S. (1971). *A net structure for semantic information storage, deduction and retrieval.* Proceedings of Second International Joint Conference on Artificial Intelligence, pp. 512–523.

Shen, Z., Ma, K. L., & Eliassi-Rad, T. (2006). Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction. *IEEE Transactions of Visualization and Computer Graphics, 12*(6), 1427-1439.

Sidner, C. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse.* (PhD Thesis), MIT, Boston, MA.

Simon, H. (1955). On a Class of Skew Distribution Functions. *Biometrika, 42*(3-4), 425-440.

Skillicorn, D. (2008). *Knowledge Discovery for Counterterrorism and Law Enforcement*. Boca Raton, FL: CRC Press.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American society for information science, 24*(4), 265-269.

Smith, A. E., & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior Research Methods, 38*(2), 262-279.

Snijders, T. (2001). The statistical evaluation of social network dynamics. *Sociological methodology, 31*, 361-395.

Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Sowa, J. F. (1992). Semantic Networks. In S. C. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* (2nd ed., pp. 1493 - 1511). New York, NY: Wiley.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004, August). *Probabilistic author-topic models for information discovery.* Proceedings of 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Seattle, WA, pp. 306-315.

Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E., Buttler, D., & Hysom, D.). Reconcile: A coreference resolution research platform. Presentation at

Stuetzer, C. M., Carley, K. M., Koehler, T., & Thiem, G. (2011, June). *The communication infrastructure during the learning process in web based collaborative learning systems.* Proceedings of 3rd International Conference on WebScience (WebSci), Koblenz, Germany.

Trigg, R., & Weiser, M. (1986). TEXTNET: a network-based approach to text handling. *ACM Transactions on Information Systems (TOIS), 4*(1), 1-23.

Tushman, M. L. (1977). Special boundary roles in the innovation process. *Administrative Science Quarterly, 22*(4), 587-605.

Van Atteveldt, W. (2008). *Semantic network analysis: Techniques for extracting, representing, and querying media content*. Charleston, SC: BookSurge.

van Cuilenburg, J., Kleinnijenhuis, J., & de Ridder, J. (1986). A Theory of Evaluative Discourse: Towards a Graph Theory of Journalistic Texts. *European Journal of Communication, 1*(1), 65-96.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., . . . Moschitti, A. (2008, May). *BART: A modular toolkit for coreference resolution.* Proceedings of 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, pp. 9-12.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research policy, 34*(10), 1608-1618.

Walker, C., Strassel, S., Medero, J., & Maeda, K. (2006). *ACE 2005 Multilingual Training Corpus. LDC2006T06*. Linguistic Data Consortium, Philadelphia.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge, New York: Cambridge University Press.

Watts, D. J. (2007). The Accidental Influentials. *Harvard Business Review, 85*(2), 22-23.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

Weaver, W., & Shannon, C. E. (1949). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Weil, S. A., Foster, P., Freeman, J., Carley, K. M., Diesner, J., Franz, T., . . . Gorman, J. C. (2008). Converging Approaches to Automated Communications-based Assessment of Team Situation Awareness. In M. P. Letsky, N. W. Warner, S. M. Flore & C. A. P. Smith (Eds.), *Macrocognition in Teams. Theories and Methodologies* (pp. 277 - 304). Aldershot: Ashgate.

Weischedel, R., & Brunstein, A. (2005). *BBN Pronoun Coreference and Entity Type Corpus. LDC2005T33*. Linguistic Data Consortium, Philadelphia.

Weischedel, R., Pradhan, S., Ramshaw, L., Kaufman, J., Franchini, M., El-Bachouti, M., . . . Taylor, A. (2011). *OntoNotes Release 4.0. LDC2011T03*. Linguistic Data Consortium, Philadelphia.

Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of social structure, 8*(2), 564-586.

White, D. R., & Reitz, K. P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks, 5*(2), 193-234.

White, H. C. (1963). *An anatomy of kinship: mathematical models for structures of cumulated roles*. Englewood Cliffs, NJ: Prentice-Hall.

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology, 81*(4), 730-780.

Whitelaw, C., Patrick, J., & Herke-Couchman, M. (2006). Identifying interpersonal distance using systemic features. In J. G. Shanahan, Q. Yan & J. M. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 199-214). Dordrecht: Springer.

Winship, C. (1988). Thoughts about roles and relations: an old document revisited. *Social Networks, 10*(3), 209-231.

Woelfel, J., Holmes, R., Cody, M., & Fink, E. L. (1988). A multidimensional scaling based procedure for designing persuasive messages and measuring their effects. In G. A. Barnett & J. Woelfel (Eds.), *Readings in the Galileo system: Theory, methods, and applications* (pp. 313-332). Dubuque, IA: Kendall-Hunt.

Woods, W. (1975). What's in a link: Foundations for semantic networks. In D. Bobrow & A. Collins (Eds.), *Representation and Understanding: Studies in Cognitive Science* (pp. 35-82). New York, NY: Academic Press.

Yang, Y., & Pedersen, J. (1997, July). *A comparative study on feature selection in text categorization.* Proceedings of 14th International Conference on Machine Learning (ICML), Nashville, TN, pp. 412-420.

Zagorecki, A., Ko, K., & Comfort, L. K. Interorganizational Information Exchange and Efficiency: Organizational Performance in Emergency Environments. *Journal of Artificial Societies and Social Simulation, 13*(3).

Zelenko, D., Aone, C., & Richardella, A. (2003). Kernel methods for relation extraction. *The Journal of Machine Learning Research, 3*, 1083-1106.

Appendix

Table 153: Full name and LDC ID number for datasets

Short name	Full name	LDC ID number
MUC 6	Message Understanding Conference (MUC) 6	LDC2003T13
MUC 7	Message Understanding Conference (MUC) 7	LDC2001T02
ACE 2	Automated Content Extraction (ACE)-2 Version 1.0	LDC2003T11
TIDES 2003	TIDES Extraction (ACE) 2003 Multilingual Training Data	LDC2004T09
ACE 2004	ACE 2004 Multilingual Training Corpus	LDC2005T09
ACE 2005	ACE 2005 Multilingual Training Corpus	LDC2006T06
reACE	Datasets for Generic Relation Extraction (reACE)	LDC2011T08
BBN	BBN Pronoun Coreference and Entity Type Corpus	LDC2005T33
SemEval 2010-8	SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals	n.a.
Onto Notes 4	OntoNotes Release 4.0	LDC2011T03
SemEval 2010-1	SemEval-2010 Task 1: OntoNotes: Coreference resolution in multiple languages.	LDC2011T01
NYT AC	The New York Times Annotated Corpus	LDC2008T19
CoNLL 2003	CoNLL-2003 task: Language-Independent Named Entity Recognition	n.a.

Table 154: Network Analysis Measures used in thesis*

Metric	Definition	Range of output values**	Input converted to	Level of analysis	Reference
Average Distance	The average shortest path length between nodes, excluding infinite distances.	0, N	square, binary	Graph	(Wasserman & Faust, 1994)
Average Speed	The average inverse geodesic distance between all node pairs. The highest score is achieved for a clique, and the lowest for all isolates	0,1	square, binary	Graph	(Carley, 2002b)
Betweenness Centrality	Per node *i*, across all node pairs that have a shortest path containing *i*, the percentage that pass through *i*.	0,1	square, binary	Node	(Freeman, 1979)
Betweenness Centralization	Network centralization based on the betweenness score for each node in a square network.	0,1	square, binary	Graph	(Freeman, 1979)
Clique Count	The number of distinct cliques to which each node in a network belongs. A clique is a maximal complete subgraph of three or more nodes.	0, N	square, symmetric	Node	(Wasserman & Faust, 1994)

Component Count Strong	The number of strongly connected components in a directed network. This is computed directly on G, whether or not G is directed.	0,N	square, binary	Graph	(Wasserman & Faust, 1994)
Component Count Weak	The number of weakly connected components in a directed network. Such components are called "weak" because the graph G is undirected.	0,N	square, binary, symmetric	Graph	(Wasserman & Faust, 1994)
Degree Centrality	The normalized in-degree plus out-degree of a node. I.e. the size of the immediate ego-network of a node.	0,1	square	Node	(Wasserman & Faust, 1994)
Degree Centralization	A centralization of a square network based on total degree centrality of each node.	0,1	square, symmetric	Graph	(Freeman, 1979)
Connectedness	Measures the degree to which a square network's underlying (undirected) network is connected.	0,1	square, symmetric	Graph	(Krackhardt, 1994)
Density	The ratio of the number of edges versus the maximum possible edges for a network.	0,1	N, L	Graph	(Wasserman & Faust, 1994)
Diffusion	The degree to which something could be easily diffused (spread) throughout the network. This is based on the distance between nodes. A large diffusion value means that nodes are close to each other, and a smaller diffusion value means that nodes are farther apart.	0,1	square, binary	Graph	(Carley, 2002b)
Efficiency	The degree to which each component in a network contains the minimum edges possible to keep it connected.	0,1	square, binary, symmetric	Graph	(Krackhardt, 1994)
Eigenvector Centrality	The centrality of a node based on its degree and the degrees of its neighbors.	0,1	square, symmetric	Node	(Bonacich, 1987)
Eigenvector Centrality	Calculates the eigenvector of the largest positive eigenvalue of the adjacency matrix representation of a square network.	0,1	square, symmetric	Graph	(Bonacich, 1987)
Fragmentation	The proportion of nodes in a network that are disconnected.	0,1	square, binary,	Graph	(Borgatti, 2003)

			symmetric		
Global Efficiency	Global Efficiency is the normalized sum of the inverse geodesic distances between all node pairs.	0,1	square, binary, symmetric	Graph	(Latora & Marchiori, 2001)
Hierarchy	The degree to which a network exhibits a pure hierarchical structure.	0,1	square, binary	Graph	(Krackhardt, 1994)
Inverse Closeness Centralization	The average closeness of a node to the other nodes in a network. Inverse Closeness is the sum of the inverse distances between a node and all other nodes.	0,1	square, binary	Graph	(Wasserman & Faust, 1994)
Network Levels	The Network Level of a square network is the maximum Node Level of its nodes. This measure is also called diameter.	0, |N|-1	square, binary	Graph	(Carley, Reminga, et al., 2011)
Clustering Coefficient	Measures the degree of clustering in a network by averaging the clustering coefficient of each node. The clustering coefficient of a node is the density of its ego network - the sub graph induced by its immediate neighbors.	0,1	square, binary	Graph	(Watts & Strogatz, 1998)
Transitivity	The percentage of edge pairs (i,j), (j,k) in the network such that (i,k) is also an edge in the network.	0,1	square, binary	Graph	(Carley, Reminga, et al., 2011)
Upper boundedness	The degree to which pairs of agents have a common ancestor.	0,1	square, binary	Graph	(Krackhardt, 1994)

* For more details on these metrics see (Carley, 2002b; Carley, Reminga, et al., 2011) . Definitions are partially preprinted from that source.

** Definitions: N = number of nodes, L = number of links

Table 155: Error Analysis, Class Model 3, absolute values

next page

Table 156: Error Analysis, Class Model 4, absolute values

two pages ahead

	Prediction																																
Ground Truth	agent na	attribute age	attribute numerical	event na	event war	knowledge art	knowledge language	knowledge law	location city	location country	location facility	location other	location state-prov	none	org. corporate	org. edu	org. gov	org. other	org. political	org. religious	org-att nationality	org-att other	org-att political	org-att religious	resource animal	resource disease	resource money	resource plant	resource product	resource substance	task game	time na	Sum
agent na	45,418		10	17		166			318	8	29	15	24	2,548	791	12	35	49	2	1	5		20		5		5		21	1	1	27	49,528
attribute age		764	160			1					1			111													1		4			52	1,094
attribute numerical	8	129	28,995			11					6	1		1,439	23			1									58		6			314	30,991
event na	18		1	426		35			8	3	4	5		68	14		3		1										4	2		37	629
event war				2	110					3				2	3						2												122
knowledge art	268		11	9		1,040		4	65	6	12	10	5	479	167	9	18	23		6	8			3	4	1			11	4		38	2,201
knowledge language	12					4	43	2				2		13							10												86
knowledge law	7		23	3	2	15		695	5	3				87	37		4	11											5			10	907
location city	435		1			58			6,667	28	27	75	17	152	335	14	20	39	4	2	3						2		1			9	7,889
location country	32		1			7		1	21	6,329	1	46	13	120	49	1	64	11				4					1						6,701
location facility	88	2	24	2		66		3	71	3	2,182	14	12	763	151	17	28	24	2		3						2		3	3		10	3,473
location other	86		8	5	1	28			101	52	19	1,475	16	147	110	2	8	12			5				2					1		5	2,083
location state-prov	51					5			102	17	1	9	2,838	430	49	2	14	4	2													6	3,530
none	1,012	52	1,489	20		517	2	40	144	25	353	60	67	889,453	2,522	28	246	207	5	7	26	3	5	2	31	34	541	12	267	255	25	3,118	900,568
org. corporate	2,161		69	3		254		12	384	134	94	182	125	4,486	54,724	37	325	252	1	1	30		1	4	1		6		49	2		45	63,382
org. edu	33					18			24	1	3	3	7	96	52	971	21	14		2									1				1,246
org. gov	112		2	5		33		10	20	23	9	31	6	375	506	3	9,691	75	3		7		1				2		8	1		2	10,925
org. other	164		7			75		8	107	10	18	48	8	478	512	16	95	3,077	3	20	5	3	1		4		3		1	1		5	4,669
org. political	37					2								94	60	2	3	65	504			1	28									2	798
org. religious	13					2			2		3			30	13	1	2	4	1	77				2					2				152
org-att nationality	88		1			7	1		13	4		6		42	32		10	4	3		3,300	7	6	2					6			6	3,538
org-att other	15						1		8	2		1		19	4			4			8	33										1	96
org-att political	32											1		11	4		5		26		2		601										682
org-att religious	6								1					6	10				1	1	7		1	56	4							1	94
resource animal	26					12			1					202	3										168				1				413
resource disease	10								1					170	1											194				1		1	378
resource money	2	2	139						1	2				607	2			3			3						30,905					20	31,686
resource plant	9								2		28			61	1													96		1			198
resource product	190	5	89	9		138		5	36	1	2	11	4	505	225	8	5	14			15				2	5	7		1,334	8		45	2,663
resource substance	42	3	31	1		2			12					931	48	1		3			3			2	2	1	1	2	20	1,697		6	2,808
task game	10													59	5																24		98
time na	20	17	532	8		17		4	3	1		3		2,165	13						9						18		2	1		39,439	42,252
Sum	50,405	974	31,593	510	113	2,513	47	784	8,117	6,655	2,792	1,998	3,142	906,149	60,466	1,124	10,597	3,896	558	117	3,451	51	664	71	223	235	31,552	110	1,746	1,978	50	43,199	1,175,880

	Predictions																																													
Ground Truth	agent_generic_na	agent_specific_na	attribute_na_age	attribute_na_numerical	event_specific_na	event_specific_war	knowledge_specific_art	knowledge_specific_language	knowledge_specific_law	location_generic_city	location_generic_country	location_generic_facility	location_generic_other	location_generic_state-prov	location_specific_city	location_specific_country	location_specific_facility	location_specific_other	location_specific_state-prov	none	org._generic_corporate	org._generic_edu	org._generic_gov	org._generic_other	org._generic_political	org._generic_religious	org._specific_corporate	org._specific_edu	org._specific_gov	org._specific_other	org._specific_political	org._specific_religious	org-att_specific_nationality	org-att_specific_other	org-att_specific_political	org-att_specific_religious	resource_generic_product	resource_na_animal	resource_na_disease	resource_na_money	resource_na_plant	resource_na_substance	resource_specific_product	task_na_game	time_na_na	Sum
agent_generic_na	25,263	57	1	5	5		32			1	4	4		1	10		2	2	9	2,224	233		10	38	2	1	41	6	16	17	1	1	2		2		3	6	1			1	4	1	7	28,013
agent_specific_na	31	19,849		8	12		151		6						290	6	30	12	24	442	5						531	14	8	17		1	1		2			3	3	1		1	47		20	21,515
attribute_na_age			732	178													2			110							1	1															6		64	1,094
attribute_na_numerical	1	4	124	28,960	3		6		1								7	2		1,446	1						20			2										72			9		333	30,991
event_specific_na		15		1	434		28		2						7	1	9	3		63							19		5		1											1	4		36	629
event_specific_war					2	113										3				2													2													122
knowledge_specific_art	30	179		13	4		1,068	2	7						67	3	39	10	3	476	1			1			169	4	11	27		7	2	2		2	1	4	5	1		4	26		33	2,201
knowledge_specific_language		8		1			3	43									1	2		15							1						12													86
knowledge_specific_law	8	7	1	29	3	1	13		703						4	1				81							18		8	14													8		8	907
location_generic_city	4									293		7	2		3			1		63	3		1																							377
location_generic_country	12									3	865	3		8						82	13		3	2			2																			993
location_generic_facility	2	2					1			1	1	1,760			4		6			753	35		6	5			5	1	1		1							1		1		3			4	2,593
location_generic_other	2		2							4		4	13					1		25	1			17																						69
location_generic_state-prov	1									2	5	1		210		1			2	172							1	1																	1	397
location_specific_city	5	303		1	1		46		2	3					6,615	27	11	69	6	136							215	13	3	34	4	3	1							1			5		8	7,512
location_specific_country	1	16		1		1	4		1					1	22	5,538		11	12	29	1						23		36	6				2				1		2						5,708
location_specific_facility	3	72	1	20	2		32		3			7			62	5	438	14	10	50			1				99	11	14	10		5		2						2			10		7	880
location_specific_other	5	67	1	6			24		1		1	2	1		95	46	8	1,447	10	127							131		10	17		2	4						1			1	3		4	2,014
location_specific_state-prov	3	33		2			2							2	139	7	1	11	2,675	193							49	1	8	3									1						3	3,133
none	635	242	62	1,481	23		516	2	40	34	9	260	2	55	101	19	38	54	81	889,849	887	11	33	65	2	5	1,371	46	160	133	3	8	20	6	3	4	196	37	40	545	11	280	58	28	3,113	900,568
org._generic_corporate	410	5	1	14			3					50					1	2		928	13,581	2	149	103	2	1	33	3	2				1				6			1	1	1			5	15,305
org._generic_edu	6										2	1	1							28	15	178	2	2	1		3	4		2																245
org._generic_gov	57									1	2	3		6	1					109	259		2,051	24	1		1		5								1									2,521
org._generic_other	72		1				4				2	20	4			1				202	157		21	827		2		1	4	18							3			2					2	1,343
org._generic_political	11										1									20	13	1	4	27	73		1																			151
org._generic_religious	1											1								16	2	2	2	1		24						1					1									51
org._specific_corporate	87	1,315	1	55	2		201		21		1			1	422	85	66	176	51	3,104	22				3		41,938	44	190	140	5	3	17	2			1	1		5		3	74		41	48,077
org._specific_edu	5	18		1			9								20	2	7	2	8	52							42	779	13	34	9															1,001
org._specific_gov	22	31		3	3		26		9					2	16	16	23	23	5	265	5		3	1	3		247	4	7,629	47	2	2	5		1		1			1		1	6		2	8,404
org._specific_other	36	59		6	1		84		12						98	5	14	36	10	249	6		1	34	1		308	17	83	2,233		12	5	4		2		2		1			2		5	3,326
org._specific_political	6	15					3								4					81				11			18	8	4	51	413				29								2		2	647
org._specific_religious	3	6					8								5		6			9			1				4		3	2		49				3							2			101
org-att_specific_nationality	27	22		6			8	3			1				16	1	1	9		62							30	1	2	4	3		3,319	8	4	2							6		3	3,538
org-att_specific_other	2	4			1		2	1							7	1		1		25							5	1		4			8	33									1			96
org-att_specific_political	17	2													1			1		13							6		2		21		2		617											682
org-att_specific_religious	13	2																		8										3		1	3		2	56		4					2			94
resource_generic_product	2	1	2				3					2					1			344	26						3	1	2								1,001		2			1	2		4	1,397
resource_na_animal	5	14					17								1					199	2			1			5											167	1				1			413
resource_na_disease	1	8													1					173							2												192			1				378
resource_na_money	1	2	2	137											2	4			2	550	4			1			1						2							30,958			2		18	31,686
resource_na_plant	1	1		2			1					28								67										2											96					198
resource_na_substance	12	13	2	34			4								8					914		1					31						1					5	2		2	1,742	31		6	2,808
resource_specific_product	17	111	6	88	4		158		5						44		26	7	4	159							197	3	6	12	3		2						3	7		7	354		43	1,266
task_na_game		1					2													60							5														1			29		98
time_na_na	5	14	21	488	7	1	31		3		2	2			2		2	1		2,158				1			13		1	1			10							20		1	3	1	39,464	42,252
Sum	26,825	22,498	960	31,540	507	116	2,490	51	816	342	896	2,155	23	286	8,067	5,772	739	1,897	2,912	906,133	15,272	195	2,288	1,161	88	33	45,589	964	8,226	2,833	466	95	3,419	59	660	69	1,214	231	251	31,620	111	2,048	668	59	43,236	1,175,880

317

I. Guideline for adding content nodes to existing networks in ORA

1. Generate a network per group (analysis -> generate reports -> characterize groups and networks -> locate sub-groups). These networks are a default output from grouping nodes.
2. Check if the node class "knowledge" already exists. If not, create one (add new node class -> knowledge).
3. In the node class editor, enter the ID and title for each node, .e.g. "transportation". The same token will serve as ID and title. This information can also be imported with the import wizard from a .csv file, which contains one header row ("knowledge"), and the content of each knowledge node in a separate line.
4. Check if a knowledge x group network already exists. If not, create one (add blank network -> source node class: groups, target node class: knowledge).
5. In the "Editor" for the knowledge x group network, connect knowledge nodes to groups by checking the respective boxes.